How To Fix The World Economic Forum – According To OpenAI GPT-5

Douglas C. Youvan

doug@youvan.com

www.youvan.ai

October 18, 2025

The World Economic Forum (WEF) convenes leaders who can influence policy trajectories, markets, and cultural narratives. That convening power is useful in a fragmented world—but it is not a mandate to decide what societies should do. This paper proposes a decision architecture that lets the WEF help, not decide: it preserves democratic authority, honors moral pluralism, and keeps adoption voluntary while still enabling coordination on shared problems. Five pillars anchor the framework. Epistemic humility requires publishing uncertainties, dissent, and error bars with every recommendation. Pluralistic grounding stages proposals within major ethical traditions—human rights, the capability approach, duty/rights in Abrahamic and Dharmic thought, and secular virtue ethics—showing points of convergence and conflict. Deliberative guardrails bind initiatives to citizen assemblies, open comment windows, red-team critiques, and impact assessments that include spiritual and cultural effects, not just GDP or CO2. Subsidiarity and consent default decisions to the most local competent level and make adoption opt-in with clear exit ramps. Transparency by design links agendas, participants, minutes, funding, conflicts of interest, and minority reports to every output. The result is a practical, testable blueprint for a WEF that clarifies options and trade-offs—while the public and its institutions retain the right to decide.

Keywords: WEF governance, decision architecture, epistemic humility, pluralistic ethics, citizen assemblies, red-team reviews, impact assessment, subsidiarity, informed consent, transparency by design, minority reports, legitimacy metrics.

A collaboration with GPT-5. 79 pages. CC4.0.

Outline

1. Problem Statement: WEF's Influence Without Mandate
   1.1 Convening vs. governing
   1.2 Democratic legitimacy risks (soft policy, agenda capture)
   1.3 Design goal: coordination without rule

2. Principles and Definitions
   2.1 Epistemic humility (claims, uncertainties, dissent)
   2.2 Pluralistic grounding (ethics matrix across traditions)
   2.3 Deliberative guardrails (citizens, comments, red teams)
   2.4 Subsidiarity & consent (opt-in, reversibility)
   2.5 Transparency by design (traceable artifacts, COI, minority reports)

3. Epistemic Humility in Practice
   3.1 Evidence ledger: data lineage, methods, confidence intervals
   3.2 Structured dissent: minority memos and mandatory responses
   3.3 Pre-committed falsifiers: "what would change our mind?"
   3.4 Uncertainty dashboards bundled with every release

4. Pluralistic Grounding: Moral Reasoning Made Visible
   4.1 Ethics matrix template: {proposal} x {frameworks}
   4.2 Worked example: cross-border carbon policy
   4.3 Managing persistent disagreement (safeguards and opt-outs)
   4.4 Faith-informed review protocols (respectful, documented)

5. Deliberative Guardrails
   5.1 Citizen assemblies / sortition: sampling, briefings, outputs
   5.2 Open drafting & public comment: versions, response logs
   5.3 Red-team reviews: adversarial stress tests and threat models
   5.4 Impact beyond economics: cultural, spiritual, dignity metrics

6. Subsidiarity & Consent
   6.1 Competence tests for "most local" decision locus
   6.2 Adoption contracts: scope, term, exit clauses
   6.3 Pilots and sunset provisions (reversibility by design)
   6.4 Respecting non-adoption: no penalty, no backdoor coercion

7. Transparency by Design
   7.1 Public artifact registry: agendas, attendees, minutes, drafts, funding
   7.2 COI/recusal schema (machine-readable)

7.3 Traceable edits and authorship diffs
7.4 Mandatory minority reports attached to outputs

8. Roles and Separation of Functions
   8.1 WEF as convener vs. analyst vs. advocate (firewalls)
   8.2 Independent ombuds and audit committee (publish-or-perish)
   8.3 Ethics rapporteurs (cross-tradition) and data stewards
   8.4 Escalation paths when principles clash

9. Metrics and Verification
   9.1 Legitimacy KPIs: participation breadth, dissent inclusion, reversal rates
   9.2 Quality KPIs: forecast calibration, replication, error-bar accuracy
   9.3 Equity & dignity indices: distributional and cultural checks
   9.4 Post-implementation scorecards and public retrospectives

10. Case Studies (Templates)
    10.1 AI safety standard under this architecture
    10.2 Climate finance mechanism with ethics matrix + citizen assembly
    10.3 Public health guidance with uncertainty dashboards + minority report

11. Implementation Roadmap for the WEF
    11.1 First 90 days: artifact registry, COI schema, red-team roster
    11.2 Six-month pilots: one initiative per pillar
    11.3 Twelve-month scale: annual legitimacy report and APIs for partners
    11.4 Interoperability with governments, NGOs, faith bodies, academia

12. Risks and Antipatterns
    12.1 Performative transparency
    12.2 Token pluralism
    12.3 Incumbent capture
    12.4 Consultation fatigue (rotate formats; compensate civic labor)

13. Ethical/Theological Appendix: Hosting Disagreement Well
    13.1 Minimal common ground (dignity, non-coercion, truthfulness)
    13.2 Conscientious objection and safe opt-outs
    13.3 Sacred-domain safeguards (speech, ritual, conscience)

14. Conclusion: Help, Don’t Decide
    14.1 What the WEF owes the public
    14.2 What publics should demand
    14.3 From aura to auditability

Appendices

- A. Ethics Matrix template (fillable)
- B. Citizen Assembly brief kit (neutral primers + question set)
- C. Red-Team playbook (threat models, checklists)
- D. Impact Assessment rubric (economic + cultural/spiritual)
- E. COI/Recusal schema (machine-readable fields)

References

1. Problem Statement: WEF's Influence Without Mandate

1.1 Convening vs. governing

What convening is:

- A platform function: bring heterogeneous actors together (states, firms, NGOs, faith and cultural institutions, scientists, labor) to exchange information, test ideas, and coordinate voluntary action.
- Non-binding output: playbooks, scenarios, best practices, pilot coalitions, MOUs.
- Legitimate scope: catalyze cooperation on cross-border problems where no single authority suffices (pandemics, climate risk, supply-chain resilience, digital standards exploration).

What governing is:

- A public authority function: set rules with coercive force, backed by law, budgets, and courts.
- Binding output: legislation, regulation, enforceable standards, fiscal transfers, sanctions.
- Legitimate scope: democratically authorized or otherwise publicly accountable institutions.

Why the distinction blurs in practice:

- Repetition and scale: widely cited "non-binding" frameworks can harden into de facto rules.
- Gatekeeping effects: who speaks and what gets published can steer agendas before parliaments ever vote.
- Signaling power: markets and ministries may treat convened "consensus" as policy guidance.

Working definition for this paper:

- *Convening = option generation and coordination without coercion.*
- *Governing = decision and enforcement with public accountability.*
  Our design problem is to keep the former from drifting into the latter.

Operational guardrail (rule of thumb):

- If adoption requires penalties, exclusive access, or regulatory compulsion, it is governing and must move to a public forum.
- If adoption is opt-in, reversible, and transparently argued with uncertainty disclosed, it fits convening.

1.2 Democratic legitimacy risks (soft policy, agenda capture)

Soft policy risk:

- *Definition:* recommendations and "voluntary" standards become de facto rules through procurement, ratings, or industry coordination—without public debate or legislative scrutiny.
- *Mechanism:* lenders/insurers/ratings agencies encode forum output into eligibility criteria; smaller actors must comply or lose access.
- *Public effect:* coercion by market infrastructure rather than by law; citizens can't contest it at the ballot box.

Agenda capture risk:

- *Definition:* issue framing tilts toward incumbent interests because agenda-setting and speaker selection are controlled by a narrow set of funders, sectors, or regions.
- *Mechanism:* topic curation, panel composition, white-paper authorship, and "invited consultations" filter dissent.
- *Public effect:* the option set shrinks; plausible alternatives are under-examined; perceived elitism and distrust rise.

Opacity risk:

- *Definition:* key artifacts (attendee lists, drafts, meeting notes, funding links) are not fully public, so influence pathways are invisible.
- *Mechanism:* private side sessions and unpublished drafts shape the "consensus" that later appears polished.
- *Public effect:* accountability gap—citizens can't see who argued what, or how trade-offs were weighed.

Representation risk:

- *Definition:* communities most affected by a proposal are underrepresented or tokenized.
- *Mechanism:* cost of attendance, language barriers, selection bias.
- *Public effect:* procedural injustice; outputs lack situated knowledge, social license, and durability.

Legitimacy diagnostics (early-warning indicators):

- Share of documents with published uncertainty/error bars < X%.
- Ratio of repeat speakers/funders to new entrants > Y.
- Fraction of sessions with structured dissent/minority reports = 0.
- % of initiatives with opt-out/exit clauses absent or unclear.
- Geographic/sectoral skew in authorship vs. affected populations.

1.3 Design goal: coordination without rule

Target outcome:
A decision architecture in which the WEF can accelerate collective problem-solving while never substituting for democratic, philosophical, or theological adjudication of ends.

Non-negotiable design constraints:

1. No coercion: all adoptions are opt-in, reversible, and clearly separable from access to essential markets or public services.
2. Traceability: every recommendation ships with sources, methods, uncertainty, funding, conflicts, and minority views.
3. Plural reasoning: proposals are mapped across major ethical frameworks; irreducible disagreement is recorded and respected.
4. Local primacy: decisions default to the most local competent level (subsidiarity); global templates become menus, not mandates.
5. Deliberative coupling: citizen assemblies, public comments, and red-team reviews are built-in, not occasional.

Design levers (what the architecture must implement):

- Evidence ledger: provenance of data/models; pre-committed "what would change our mind" criteria.
- Ethics matrix: a standard table that shows how human rights, capabilities, duties/rights (Abrahamic, Dharmic), and virtue ethics score a proposal—plus narrative rationale and objections.
- Guardrail stack: sortition panels, open drafts with response logs, adversarial stress tests, and cultural/spiritual impact assessments alongside economic ones.
- Adoption contract: a plain-language template with scope, term, review cadence, and exit ramps; explicitly states that non-adoption entails no penalties.
- Artifact registry: public, machine-readable posting of agendas, attendees, minutes, drafts, funding links, COI/recusals, and minority reports.

Success criteria (measurable):

- *Legitimacy:* broader participation, visible dissent inclusion, documented reversals when evidence changes.
- *Quality:* calibrated forecasts, replication of claims, narrowing error bars over iterations.
- *Equity & dignity:* demonstrated attention to distributional, cultural, and conscience impacts; honored opt-outs.
- *Adoptability:* voluntary uptake across diverse locales because options are clear, reversible, and respectful—not because access or status is at stake.

One-sentence design brief:
Build processes that clarify choices and coordinate willing actors—but that cannot be used to compel the unwilling or to bypass the forums where the right is properly decided.

2. Principles and Definitions

2.1 Epistemic humility (claims, uncertainties, dissent)

Definition. Epistemic humility is the practice of presenting claims together with their limits: assumptions, confidence intervals, known failure modes, and structured dissent.

Why it matters. Conveners aggregate signal but don't own truth. Humility preserves credibility and prevents "soft policy" from hardening into dogma.

Operational elements.

- Claim card. For every recommendation: (a) statement of claim, (b) evidence summary, (c) assumptions, (d) uncertainty/CI or qualitative error bars, (e) external validity (where it *doesn't* apply), (f) "what would change our mind" criteria.
- Uncertainty dashboard. Visual panel pairing point estimates with ranges; show sensitivity to key assumptions and the direction/magnitude of model risk.
- Dissent protocol. Named minority memos attached to artifacts (max 1–2 pages each) with three parts: the objection, the alternative reading, the decisive test that could resolve it.
- Replication note. Minimal materials/methods to let a third party re-run the analysis; if not possible (e.g., proprietary data), state why and what surrogate evidence exists.

Quality metrics.

- % of outputs shipping with uncertainty dashboards and dissent memos
- Forecast calibration (Brier score / calibration plots) for time-stamped predictions
- Rate of post-release corrections and reversals (with reasons)

2.2 Pluralistic grounding (ethics matrix across traditions)

Definition. Pluralistic grounding makes moral reasoning explicit by mapping each proposal across multiple ethical frameworks and living traditions, documenting convergences, conflicts, and non-negotiables.

Ethics Matrix

Framework	Verdict	Core Lens	Benefits / Harms (≤1–2 lines)	Conditions / Safeguards
Human rights	[Y/N/C]	Rights, due process	[Who benefits/loses; discrimination risk]	[Due process; grievance path]
Capabilities	[Y/N/C]	Real freedoms	[Capabilities expanded/at risk]	[Backstops; parity access]
Virtue ethics	[Y/N/C]	Character, prudence	[Virtues fostered; vices risked]	[Pilot; oversight]
Abrahamic	[Y/N/C]	Stewardship, justice	[Care for poor; truthful measures]	[Exemptions; transparency]
Dharmic	[Y/N/C]	Dharma, ahimsa	[Harm reduction; intention]	[Non-violent transitions]
Secular pluralism	[Y/N/C]	Consent, neutrality	[Coercion/neutrality check]	[Opt-in; exit ramps]

Process notes.

- Authentic voice. Use designated *ethics rapporteurs* or councils drawn from each tradition; require sourceable reasoning (texts, doctrines, contemporary scholarship).
- Handling disagreement. Where verdicts diverge, add *safeguard clauses* (opt-outs, conscientious objection, domain exclusions).
- Narrative layer. One short paragraph per row explaining *why* that framework lands where it does; avoid flattened checklists.

Success indicators.

- Diversity of traditions engaged and cited
- Presence of safeguard clauses where divergence is irreducible
- Public comprehension (readability tests) of ethics matrices by non-experts

---

2.3 Deliberative guardrails (citizens, comments, red teams)

Definition. Guardrails are coupled processes that ensure proposals are stress-tested by ordinary citizens, open commenters, and adversarial reviewers before promotion.

Components.

- Citizen assemblies (sortition). Randomly selected, demographically stratified participants receive neutral primers, question lists, and facilitated sessions; outputs = recommendations plus minority views.
    - *Protocol:* sample frame → conflict-of-interest screening → balanced briefings → Q&A with proponents/opponents → deliberation → written output.
- Open drafting & public comment. Versioned drafts posted with 30–60 day comment windows; maintain a response log indicating how each substantive comment was handled (accepted, modified, rejected with reasons).
- Red-team reviews. Independent teams tasked to break assumptions, model perverse incentives, and explore failure cascades, including cultural and spiritual harms—not only technical or economic ones.

Instruments.

- Facilitation rubric (neutrality rules, speaking time, expert rotation)
- Threat-model checklist (misuse, capture, distributional harms, sacred-domain conflicts)
- Comment triage schema (duplicate, out-of-scope, substantive, systemic)

KPIs.

- Assembly participation breadth and satisfaction
- % of substantive comments addressed in the response log
- Number of red-team findings leading to material changes or safeguards

---

2.4 Subsidiarity & consent (opt-in, reversibility)

Definition. Decisions should be taken by the most local competent level; uptake must be voluntary, reviewable, and reversible without penalty.

Adoption contract (plain-language template).

- Scope & locus. What domain, where, and by whom (local/community/sector)?

- Term & cadence. Trial period, review intervals, sunset date.
- Exit ramps. How to opt out at any time; how to revert; no retaliation clauses.
- Interface duties. What data to report back (for learning) and what privacy limits apply.
- Non-adoption neutrality. Explicit clause that declining carries no loss of essential access or status.

Competence test (choose the locus).

- Knowledge proximity (local context vs. global externalities)
- Capacity to implement and monitor
- Stakes and rights at issue (if sacred/conscience is engaged, push decisions closer to the affected community)

Design signals of real consent.

- Multiple viable options (menu, not mandate)
- Comparable support/resources across options
- Independent advice available to adopters and non-adopters

Measures.

- Fraction of pilots terminated or modified via exit ramps
- Geographic/sectoral diversity in uptake patterns (evidence of choice, not herding)
- Documented cases of conscientious objection respected without penalty

---

2.5 Transparency by design (traceable artifacts, COI, minority reports)

Definition. Every step—from agenda-setting to final outputs—must be publicly traceable, with machine-readable artifacts, conflict-of-interest (COI) disclosures, and preserved minority reports.

Artifact registry (public & API-served).

- For each initiative: agenda, participants (roles/affiliations), preparatory materials, draft versions with diffs, funding sources, COI/recusal statements, meeting minutes, uncertainty dashboard, ethics matrix, citizen assembly output, red-team report, public

comments & response log, final output, minority reports, and post-implementation reviews.

- Formats: human-readable HTML/PDF + machine-readable JSON; stable permalinks and hashes for version integrity.

COI/recusal schema.

- Relationship type (financial, advisory, familial, institutional)
- Magnitude banding (e.g., <$5k, $5–50k, >$50k; time-bound)
- Recusal triggers (topics, votes, drafting control)
- Auto-flag logic in registry (e.g., when author drafts in a domain of declared interest)

Minority reports.

- Guaranteed slot and equal permanence in the registry
- Citation parity in the final artifact (explicit references within the main text)
- Follow-up hooks (conditions under which the minority view should be re-evaluated)

Anti-opacity practices.

- Publish attendance and side-session logs
- Prohibit unverifiable “consensus” claims; instead publish consensus contours (what’s agreed, contested, unknown)
- Postmortems for any initiative that skips a required artifact (with reasons)

Audit metrics.

- Registry completeness score per initiative
- COI disclosure rate and timeliness
- Minority-report inclusion and later impact (e.g., triggered revisions)

Section takeaway.
These five principles are not slogans; they are *interfaces*. Each principle comes with a concrete artifact, a protocol, and measurable KPIs so that convening power remains helpful—never substitutive of democratic, philosophical, or theological authority.

3. Epistemic Humility in Practice

3.1 Evidence ledger: data lineage, methods, confidence intervals

Purpose. Make every empirical claim traceable from dataset to conclusion so outside reviewers can re-run, stress-test, and contest it.

Artifact: Evidence Ledger (machine-readable + human-readable).

- Provenance.
    - dataset_id, owner, license, acquisition_date, collection_method
    - sampling frame, inclusion/exclusion criteria
    - known biases (coverage, nonresponse, survivorship), mitigation steps
- Transformations.
    - preprocessing pipeline (deduping, imputation rules, outlier handling)
    - feature engineering (definitions, units, scaling)
    - audit trail with hash of each step (v1.2.0 -> v1.2.1 etc.)
- Methods.
    - model family and justification (e.g., GLM, random forest, causal DAG)
    - identification strategy (if causal), assumptions list
    - hyperparameters and cross-validation scheme
- Estimation outputs.
    - point estimates with 95% CIs or credible intervals
    - effect sizes in natural units and standardized units
    - sensitivity checks (leave-one-out, alt priors, alt specs)
- External validity.
    - scope of inference (who/where/when)
    - failure conditions (contexts where results likely break)
- Reproducibility bundle.
    - code notebooks or pseudo-code, seed values, container fingerprint

- data access instructions or synthetic surrogate if restricted

Minimal header (JSON example).


```
{
  "claim_id": "AI-Std-2025-03",
  "datasets": [
    {"id":"ECDC-Respiratory-2024Q4","license":"CC-BY-4.0","collection_method":"sentinel clinics","biases":["urban-overrep"],"mitigation":"poststratification"}
  ],
  "methods":{"model":"logit","cv":"5-fold","causal_id":"front-door","assumptions":["ignorability_cond1","stable_unit_treat"]},
  "estimates":{"or":1.42,"ci_95":[1.25,1.61]},
  "sensitivity":{"spec_changes_that_flip_sign":0,"leverage_points":3},
  "external_validity":{"geos":["EU-27"],"time_window":"2024-10..2025-03","failures":["low-reporting regions"]},
  "repro":{"container_sha":"c2ab…","seed":1729}
}
```


Quality gates.

- No release without a filled ledger.
- Ledger diffs auto-published with each revision.
- CI/credible intervals required; if impossible, supply qualitative uncertainty with rationale.

KPIs.

- % of claims with complete ledgers
- Replication rate within 90 days
- Number of ledger-prompted corrections per year

3.2 Structured dissent: minority memos and mandatory responses

Purpose. Move disagreement from rumor to record. Dissent improves truth-finding and legitimacy when it is visible, specific, and answered.

Artifact: Minority Memo (max 2 pages, attached to the main artifact).

- Scope. Which claim_id or section?
- Objection. Concrete point of disagreement (method, data, ethics).
- Counter-analysis. Alternative model or reading (minimal viable evidence).
- Test. Specific observation that would adjudicate the dispute.
- Conditions. If proposal proceeds, what safeguards are required?

Process.

1. Windowed intake. 10 business days after draft publication; late memos allowed with justification.
2. Triage. Substantive vs. non-substantive; duplicates merged.
3. Mandatory response. Authors must issue a point-by-point reply (accept/modify/reject) with reasons and, when relevant, code/analysis updates.
4. Parity in permanence. Memos and responses live alongside the final artifact, version-locked.

Response log (table fields).

- memo_id, status (accepted/modified/rejected), rationale, resulting change (y/n), link to diff, reviewer name/affiliation

Guardrails.

- No personal attacks; memos must address claims or methods.
- COI disclosure for memo authors.

KPIs.

- % of drafts receiving at least one minority memo
- Median response time to memos
- Share of memos that trigger material changes or safeguards

3.3 Pre-committed falsifiers: “what would change our mind?”

Purpose. Prevent motivated reasoning by declaring, in advance, the observations that would downgrade confidence or reverse a recommendation.

Artifact: Falsification Sheet (bundled with each claim).

- Primary falsifiers. Concrete, observable conditions that, if met, reduce confidence below a threshold (e.g., effect size shrinks by >50% in two independent replications; calibration error > 0.1 Brier in Q2).
- Secondary falsifiers. Early warning anomalies (model drift, distribution shift, harm reports) that trigger review.
- Decision thresholds. Pre-specified actions: revise, pause, withdraw.
- Monitoring plan. Who watches which signals, how often, with public reports.
- Sunset clause. If monitoring cannot be maintained, recommendation expires.

Examples (illustrative).

- *AI auditing guideline:* If two accredited labs fail to reproduce claimed robustness gains within 60 days using the evidence ledger, downgrade recommendation from “adopt” to “pilot-only.”
- *Climate finance mechanism:* If administrative overhead exceeds 12% in three consecutive quarters, pause expansion and require design changes.

KPIs.

- % of artifacts with falsifiers and thresholds defined
- Number of triggered reviews vs. silent drift discoveries
- Time from trigger to public update

3.4 Uncertainty dashboards bundled with every release

Purpose. Make uncertainty a first-class object for readers: not an afterthought buried in appendices, but a front-page companion to each recommendation.

Dashboard components (single page, consistent layout).

- Estimate + interval. Point estimate with 50% and 95% intervals; show both to avoid false precision.

- Sensitivity tiles. How the estimate moves under key assumption toggles (A/B priors, alt model family, exclusion of top-k leverage points).
- Forecast calibration. Reliability diagram for any probabilistic claims; recent Brier/Log score.
- Scenario bands. Best-case, base-case, worst-plausible paths with triggers.
- External validity map. Heatmap of contexts where evidence is thin or non-applicable.
- Dissent summary. Count of minority memos; brief of top 2 objections with links.
- Falsifier panel. The pre-committed triggers and their current status (green/amber/red).

Design rules.

- Plain units; avoid spurious decimals.
- Color used sparingly and consistently (e.g., green/amber/red for falsifier status).
- Tooltips for definitions; downloadable CSV/JSON of the underlying numbers.

Update policy.

- Dashboard version increments with any material change to data or methods.
- Archived dashboards remain linkable; changelog shows what shifted and why.

KPIs.

- Reader comprehension scores from quick surveys (e.g., can a lay reader identify the main uncertainty drivers?)
- Frequency of media/policy citations that include intervals (proxy for uptake of uncertainty norms)
- Share of releases where worst-plausible scenario triggered preparatory safeguards

---

Section takeaway.
Humility is operationalized through four mandatory artifacts: an evidence ledger, minority memos with responses, pre-committed falsifiers, and an uncertainty dashboard. Together they create a culture where recommendations are auditable, contestable, and revisable—and therefore more trustworthy.

4. Pluralistic Grounding: Moral Reasoning Made Visible

4.1 Ethics matrix template: {proposal} x {frameworks} with normative notes

Purpose. Make value judgments explicit and comparable, so readers can see *why* a proposal is endorsed, opposed, or conditionally supported across traditions.

Filling rules.

- Use Y = endorse, N = oppose, Cond = conditional/with safeguards.
- Every cell under “Normative notes” must cite a principle or text (e.g., UDHR Art. 7; Qur’an 4:135; Nicomachean Ethics II; Nussbaum’s central capabilities list).
- If a tradition is internally diverse, specify which school/text basis you used.
- Keep “Conditions/safeguards” concrete (e.g., opt-out for conscientious objectors; compensatory justice fund; data minimization clause).

Traceability. Each matrix references a short bibliography; rapporteurs sign off on their row and disclose conflicts of interest.

4.2 Worked example: cross-border carbon policy (rights, duties, capabilities, virtues)

Proposal (summary). A voluntary cross-border carbon adjustment (CBCA) mechanism where jurisdictions that price carbon can levy a border adjustment on imports from jurisdictions without equivalent measures, with: (a) exemptions for least-developed countries (LDCs), (b) revenue recycling to domestic households and an international adaptation fund, (c) an opt-in technical assistance track to help exporters decarbonize.

Ethics Matrix (template; one row per framework)

Framework	Verdict (Y/N/C)	Rationale (≤2 lines)	Conditions / Safeguards
Human rights (UDHR/ICCPR/ICESCR)			
Capability approach (Sen/Nussbaum)			
Virtue ethics (Aristotle/MacIntyre)			
Abrahamic duty/rights			
Dharmic traditions			
Secular pluralism/liberalism			

Narrative notes (illustrative, 1–2 per framework).

- *Human rights:* Ensure due process in carbon-intensity determinations; publish appeals path; audit for indirect discrimination against small exporters.
- *Capabilities:* Pair the CBCA with cash transfers and retraining to protect the income/employment capabilities of at-risk groups.
- *Virtues:* Favor phronesis (practical wisdom) through small pilots; beware signaling games that cultivate vanity rather than temperance.
- *Abrahamic:* Stewardship (Genesis 2:15), justice for the poor (Prov 14:31), and truthful measures (Lev 19:35-36) endorse action when calibrated to protect the least.
- *Dharmic:* Ahimsa and right livelihood support decarbonization pathways that reduce harm without sudden, uncompassionate disruption.
- *Secular pluralism:* Preserve consent and neutrality; ensure the mechanism does not become a disguised restriction on trade.

4.3 Handling irreducible disagreement: safeguard clauses and opt-outs

When rows disagree. Treat disagreement as a design input, not a bug. Add targeted safeguards where the *reason* for dissent identifies a remediable risk.

Safeguard catalogue (pick and tailor).

- Conscience/ritual opt-outs: Explicit legal/contractual carve-outs for actors with grounded religious or ethical objections, plus non-retaliation clauses.
- Distributional backstops: Compensation floors, transition funds, or phased timelines for communities bearing concentrated costs.
- Domain exclusions: Areas where a tradition flags sacred concerns (e.g., biometric tracking at places of worship) are excluded by default unless freely and explicitly consented to by the affected community.
- Double-key triggers: High-risk provisions activate only when both a citizen assembly and an ethics council approve.
- Sunset + re-consent: Policies expire unless re-authorized after impact review and new public consent.

Decision record. For each safeguard, document: (a) the dissent it addresses, (b) the mechanism chosen, (c) the metrics that will show whether the safeguard works, (d) an exit condition if it fails.

If disagreement is foundational (not fixable). Offer a menu of options, including "do nothing now," and maintain a stable status quo until a legitimate public authority adjudicates.

4.4 Faith and culture inputs: procedures for respectful theological review

Goal. Invite authentic reasoning from living traditions without instrumentalizing them, and ensure proposals do not trample sacred domains.

Process (lightweight, repeatable).

1. Scoping. Identify potential sacred-domain touchpoints early (e.g., bodily integrity, ritual spaces, family life, symbols).
2. Rapporteur selection. For each tradition implicated, appoint 2–3 rapporteurs: one scholar/clergy, one practitioner with community standing, optionally one legal expert. Publish bios and COIs.
3. Briefing packet. Neutral summary of the proposal; specific questions; potential impacts; data summaries; timelines.
4. Dialogue window (2–4 weeks). Structured interviews or roundtables; record reasoning, cite sources (scripture, doctrine, jurisprudence, authoritative commentaries).

5. Memo format. Each rapporteur submits a 1–2 page memo: (a) verdict (Y/N/Cond), (b) reasons, (c) conscience/sacred concerns, (d) suggested safeguards/opt-outs, (e) references.
6. Synthesis. Ethics editor compiles a comparative summary; unresolved tensions are flagged, not smoothed over.
7. Feedback loop. Proposers respond in writing, noting which safeguards they adopt and why.

Respect norms.

- No “quote-mining” sacred texts; include context and recognized interpretations.
- Avoid false equivalence: note when a tradition is internally plural and present the range.
- Consent to publish: rapporteurs approve the final text that cites them.

Cultural impact check (beyond theology).

- Use a short cultural impact rubric: language accessibility; symbolism conflicts; historical traumas; indigenous rights; local customs.
- Require free, prior, and informed consent (FPIC) where indigenous lands/rights are implicated.

Verification.

- Maintain a public ledger of theological/cultural reviews linked to each proposal.
- Track whether suggested safeguards were integrated and whether any conscientious objection cases arose post-implementation.

Section takeaway.
Pluralistic grounding is not a preface; it is an *engineered interface* between proposals and the moral landscape. The ethics matrix, worked examples, safeguard clauses, and respectful theological review make trade-offs explicit, disagreement legible, and consent real—so coordination can proceed without pretending to settle what only publics and traditions can decide.

5. Deliberative Guardrails

5.1 Citizen assemblies / sortition: sampling, briefings, outputs

Purpose. Bring ordinary people—those affected by proposals—into structured, representative, and informed deliberation.

Sampling (representativeness first).

- Frame. Build a stratified sampling frame reflecting demographics (age, gender, income, education), geography, and salient lived-experience strata (e.g., farmers, gig workers, small exporters).
- Method. Two-stage selection: (1) random invitation oversampled for underrepresented groups; (2) stratified random draw among acceptances with modest honoraria and accessibility supports (childcare, travel, translation).
- Bias controls. Publish response rates by stratum; re-weight briefings and facilitation resources where acceptance skews.

Briefings (neutral, balanced, comprehensible).

- Primer pack. Plain-language overview, pro/con memos, data snapshots, uncertainty notes, and a glossary; all under 30 pages, Flesch-Kincaid ≤ 10th grade.
- Expert panels. Equal time for proponents/critics; Q&A recorded; experts disclose COI.
- Localization. Add “lived-context” inserts (case studies, indigenous/faith perspectives, small-business impacts).

Deliberation & outputs.

- Facilitation. Trained neutrals, timed turns, rotating discussant roles; minority voices protected.
- Decision artifacts. (a) Majority recommendation with rationale; (b) documented minority statements; (c) a *conditions list* (what would change minds).
- Publication. Video, transcript, datasets, and a 2-page citizen summary (what we decided, why, open questions).

Quality indicators (KPIs).

- Participation breadth (quota fill rate; dropout rate)
- Knowledge gain (pre/post quiz deltas)

- Deliberative quality (speaking-time equality, civility scores)
- Policy influence (share of assembly conditions integrated into the final design)

5.2 Open drafting & public comment: versions, response logs

Purpose. Keep proposals corrigible in public, not just in private rooms.

Versioning protocol.

- Draft ladder. v0.1 concept → v0.5 structured draft → v0.9 near-final; each stage has a fixed comment window (30/45/30 days).
- Changelogs. Human-readable summary + machine-readable JSON diff (sections touched, reasons).

Comment intake & triage.

- Channels. Web portal, email, accessible hotline; accept attachments, data, and code.
- Schema. Every comment gets a receipt_id and metadata (author type: citizen, NGO, firm, scholar; jurisdiction; COI declaration).
- Triage buckets. Duplicate, non-substantive, substantive-technical, substantive-ethical, systemic-process.

Response log (mandatory).

- Fields. receipt_id, classification, disposition (accepted/modified/rejected), rationale (≤150 words), linked diff or rebuttal memo.
- Service levels. Substantive comments must receive individualized responses before a draft can advance a version.

Anti-astroturf measures.

- Rate-limiting by organization; clustering similar submissions; transparency on "campaigned" comments while still counting unique signatories.

KPIs.

- Median response time per category
- % of substantive comments generating changes
- Readability score of the near-final draft (avoid expert-only capture)

- Share of comments from impacted communities vs. general public

---

5.3 Red-team reviews: adversarial stress tests and threat models

Purpose. Proactively surface failure modes—technical, economic, legal, and cultural/spiritual—before they happen.

Red-team charter.

- Independence. Teams selected and funded by an oversight committee (not by proposal authors); COI disclosures required.
- Mandate. Maximize constructive adversarial critique: break assumptions, simulate exploitation/capture, test distributional edge cases, and explore sacred-domain conflicts.

Threat modeling (minimum scope).

- Abuse & misalignment. How can actors game metrics or coerce adoption indirectly (ratings, procurement, gatekeeping)?
- Distributional harm. Who bears costs? Could a minority be systematically disadvantaged or dignitary harms occur (stigma, surveillance near sacred sites)?
- Institutional capture. How can incumbents lock in advantages (IP, standards committees, auditing monopolies)?
- International spillovers. Trade, sovereignty, indigenous rights, humanitarian exceptions.

Artifacts.

- Attack tree. Structured diagram of plausible exploits with likelihood × impact scoring.
- Counterfactuals. Scenarios where key assumptions invert (e.g., input data shifts, geopolitical shock).
- Safeguard pack. Concrete mitigations: rate limits, audit separation, cap-and-sunset clauses, opt-outs, FPIC requirements, double-key approvals.

Exercise format.

- Tabletop simulation with mixed stakeholders; red-team “injects” surprises.
- Live-fire tests where safe and legal (e.g., bug-bounty-like exercises for process, not just code).

KPIs.

- Number of material vulnerabilities identified and fixed pre-release
- Time-to-mitigation for high-severity issues
- Post-mortem concordance (did real failures match flagged threats?)

5.4 Impact beyond economics: cultural, spiritual, dignity metrics

Purpose. Expand evaluation beyond GDP/CO2 to include what communities often care about most: identity, conscience, meaning, and relational goods.

Rubric (scored 0–3; publish raw notes + justifications).

1. Dignity & respect. Are people treated as ends, not means? Any stigmatizing categories or compelled speech/ritual?
2. Conscience space. Are conscientious objections feasible without penalty? Are sacred domains protected by default?
3. Relational cohesion. Effects on family, neighborhood solidarity, interfaith relations, and trust.
4. Cultural continuity. Alignment with language, customs, heritage; protections for indigenous knowledge and FPIC.
5. Symbolic resonance/harms. Likely clashes with symbols, rites, or moral narratives (e.g., biometric use in sacred spaces).
6. Voice & agency. Did affected groups co-author the design? Are exit ramps real?
7. Distributive justice. Who pays/benefits across class/region/identity? Are compensations timely and dignified?

Methods.

- Mixed evidence. Rapid ethnography, participatory workshops, community surveys, and literature scans; cite sources and sample frames.
- Panel review. Cultural/theological rapporteurs rate the rubric independently; publish dispersion (where they disagree).
- Traceability. Tie each score to a paragraph-length rationale with references.

Outputs.

- Impact brief (≤5 pages). Plain-language summary + table of scores + recommended safeguards.
- Monitoring hooks. Indicators to watch post-adoption (complaint rates, opt-out usage, festival/event conflicts, places-of-worship incidents).

KPIs.

- % of initiatives with published cultural/spiritual impact briefs
- Number of safeguards adopted in response to high-risk rubric scores
- Post-adoption grievances resolved within defined SLAs (and without retaliation)

Section takeaway.
Citizen assemblies, open drafting, red-team reviews, and dignity-aware impact assessments form a guardrail stack. Together they ensure proposals are representative, corrigible, stress-tested, and humane—so conveners can accelerate coordination without substituting for democratic or spiritual authority.

6. Subsidiarity & Consent

6.1 Competence tests for "most local" decision locus

Purpose. Assign decisions to the lowest level that can act responsibly, with higher levels providing support—not override.

Three-part test (apply in order).

1. Knowledge proximity.
    - *Indicators:* local data availability; tacit know-how; cultural/linguistic fit; presence of affected communities.
    - *Heuristic:* if ≥2 of these are strongly local, default local.
2. Execution capacity.
    - *Indicators:* staffing/skills; budget; legal authority; monitoring/feedback tools.
    - *Heuristic:* if capacity is < threshold, higher levels may *lend* capability (toolkits, auditors) without taking the decision right.

3. Externalities & rights.
    - *Indicators:* cross-border spillovers; risk to fundamental rights/conscience; intergenerational harms.
    - *Heuristic:* if material cross-jurisdiction harms or fundamental-rights constraints exist, co-decision or standard-setting may occur at a higher level—but *implementation choices* should remain local.

Decision record (one page).

- Locus chosen (village/city/sector/national/transnational)
- Evidence for each test + citations
- Support pledged from higher levels (if any)
- Review date (when locus will be re-evaluated)

KPIs.

- % of initiatives with completed competence tests
- Share of decisions retained at local vs. escalated levels (with reasons)
- Number of escalations reversed after capacity-building

6.2 Adoption contracts: scope, term, exit clauses

Purpose. Make consent explicit, informed, and revocable via a plain-language “contract” between the proposer and the adopting community/organization.

Template (plain language; 2–3 pages max).

1. Scope. What exactly is being adopted (modules/features), where, and for whom.
2. Term. Start date, trial length, review cadence (e.g., quarterly), automatic sunset.
3. Performance goals. Measurable targets + acceptable ranges; “what would change our mind” triggers.
4. Data & privacy. What will be collected, minimal necessary use, retention limits, sharing rules.
5. Safeguards. From the ethics matrix and red-team pack (e.g., opt-out pathways; FPIC where indigenous rights are implicated).

6. Governance. Who can modify the contract; who audits; complaint process and service-level timelines.
7. Exit clauses. How to pause, roll back, or terminate—at any time—without penalty.
8. Non-adoption neutrality. A bold, boxed clause: *Declining or exiting this proposal will not restrict access to essential services, public funds you otherwise qualify for, or legal standing.*

Process requirements.

- Publish draft contract with a 30-day comment window and response log.
- Signatories: adopting authority + representative(s) of affected community; COI disclosures for both.

KPIs.

- % of adoptions with a signed public contract
- Median time to honor an exit request
- Number of complaints resolved within SLA; number overturned on appeal

---

6.3 Pilots and sunset provisions (reversibility by design)

Purpose. Build *time* into consent—start small, review often, and end automatically unless re-authorized.

Pilot design rules.

- Smallest responsible slice. Limit scope (users/regions/processes) to reduce harm exposure.
- Pre-registration. Publish success metrics, monitoring plan, and falsifiers before launch.
- Dual baselines. Choose both a statistical baseline (metrics) and a *dignity baseline* (no increase in complaints/rights conflicts).

Sunset provisions.

- Automatic expiry. Contract terminates unless re-adopted after impact review and fresh consent.
- Ratcheting guardrails. If harms exceed thresholds, subsequent pilots must include stronger safeguards or narrower scope.

- Archive & learn. Publish a pilot post-mortem (what worked/failed, data, dissent).

Rollback kit (must be prepared up front).

- Data un-wind (how to delete/return data safely)
- Operational reversion plan (restore prior processes within N days)
- Communication scripts for users and affected communities

KPIs.

- Fraction of initiatives launched as pilots vs. full rollouts
- Share of pilots that sunset without re-authorization (evidence reversibility is real)
- Mean time to rollback when triggers fire

6.4 Respecting non-adoption: no penalty, no backdoor coercion

Purpose. Ensure "voluntary" means *voluntary*—no hidden sticks, no access gating, no reputational blackmail.

Prohibited practices.

- Essential-service tying. Conditioning access to utilities, benefits, permits, education, or health on adoption.
- Procurement lockouts. Designing standards so only adopters can bid for unrelated public contracts.
- Ratings gatekeeping. Embedding adoption as a prerequisite in private ratings that determine market access (finance/insurance) without public oversight.
- Dark patterns. Default-on enrollment, confusing opt-outs, or shaming UX ("Are you sure you want to put others at risk?").

Detection & remedies.

- Coercion audit. Independent review of incentives, contracts, and market infrastructures; publish findings.
- Redress path. Fast-track complaints; ability to compel removal of coercive clauses; public naming of violations.

- Parity guarantees. Provide non-adopters with equivalent access to public information, forums, and baseline support.

Signaling real choice.

- Present at least two viable alternatives (including status quo), each resourced and described with pros/cons.
- Offer "trial without data capture" paths where feasible.
- Publicly report adoption *and* non-adoption rates with reasons gathered through neutral surveys.

KPIs.

- Number of validated coercion complaints per year (target → decreasing to zero)
- % of proposals offering a fully resourced "do nothing now" option
- Access parity index (non-adopters vs. adopters across key services)

Section takeaway.
Subsidiarity and consent become *operational* when you: test for the right decision locus, bind uptake to plain-language adoption contracts, default to pilots with automatic sunsets, and police coercion. The result is coordination that communities can *enter* and *exit* on their own terms—genuine consent, not compliance theater.

7. Transparency by Design

7.1 Public artifact registry: agendas, attendees, minutes, drafts, funding

Purpose. Make every step—from scoping to sunset—independently auditable by the public, press, researchers, and affected communities.

Design requirements.

- Single canonical registry with human-readable pages and machine-readable APIs (JSON/CSV); stable permalinks and content hashes for integrity.
- Full lifecycle coverage: scoping note → agenda → attendee list → prep packets → draft ladder (v0.1/0.5/0.9) → comments & responses → red-team report → citizen assembly output → ethics matrix → uncertainty dashboard → adoption contracts → minority reports → post-implementation reviews → sunset/postmortem.

Minimum fields (per initiative, initiative_id).

- Meta: title, summary, tags, jurisdictions, affected populations, start_date, current_stage, maintainers.
- Agendas: meeting_id, date/time/zone, objectives, topics, links to materials.
- Attendees: name, role, affiliation, country, COI link, recusal flags, speaking slots/time.
- Minutes: decisions, alternatives discussed, votes (if any), open questions, action items with owners/due dates.
- Drafts: version, changelog summary, full text, diff links, comment_window dates.
- Funding: sources (entity, type: grant/fee/in-kind), amounts/bands, time-bounded commitments, restrictions.
- Evidence/Methods: evidence_ledger link, datasets, model specs, code/notebook links or reason for unavailability.
- Ethics & Deliberation: ethics_matrix link, citizen_assembly report, red_team report, cultural/spiritual impact brief.
- Adoption: contract text, scope, term, exit clauses, participating entities.
- Outcomes: KPIs, calibration logs, complaints/appeals statistics, triggered falsifiers.
- Minority reports: list with authors, claims disputed, proposed safeguards, response links.
- Integrity: sha256 of each artifact, supersedes/superseded_by pointers.

API sketch.

- GET /api/v1/initiatives (query by tag/jurisdiction/stage)
- GET /api/v1/initiative/{initiative_id} (full bundle)
- GET /api/v1/initiative/{initiative_id}/drafts?v=0.9
- GET /api/v1/initiative/{initiative_id}/diff?from=0.5&to=0.9
- GET /api/v1/initiative/{initiative_id}/funding
- GET /api/v1/initiative/{initiative_id}/minority-reports

Governance.

- Publish-or-perish rule: no public launch or outreach until mandatory artifacts are posted.

- Time-to-publish SLAs: agendas ≥72h before meetings; minutes ≤10 business days; draft diffs and response logs with each version bump.

KPIs.

- Registry completeness score (share of mandatory fields populated)
- SLA adherence rate for agendas/minutes/diffs
- API reliability (>99.5% uptime) and daily unique fetches (usage proxy)

7.2 COI/recusal schema (machine-readable)

Purpose. Standardize conflict declarations and automate integrity checks.

Schema (entity = person or organization).


```
{
  "entity_id": "pers_00987",
  "name": "Jane Doe",
  "roles": ["author","panelist"],
  "affiliations": [{"org":"Acme Labs","since":"2022-03"}],
  "coi": [
   {"type":"financial","description":"Consulting","counterparty":"Acme Labs","amount_band":"$5k–$50k","period":"2024","evidence_url":"…"},
   {"type":"fiduciary","description":"Board member","counterparty":"GreenFound","unpaid":true,"period":"2023–present"}
  ],
  "topic_scope": ["AI safety metrics","carbon accounting"],
  "recusal_rules": [
{"trigger":"work_on_AI_safety_metrics","applies_to":["drafting","chairing","vote"],"start":"2024-09-01","end":"2026-09-01"}
  ],
```

"last_attested":"2025-10-12",

"attestor":"Jane Doe"

}

Operational rules.

- Granularity: banded amounts, time-bounded periods; update cadence ≥ quarterly or before relevant meetings.
- Triggers: automatic flags when counterparty intersects an initiative's funding, when topic_scope overlaps a draft's domain, or when a declared role conflicts (e.g., vendor authoring a standard they would certify).
- Recusal log: who recused, from what, by which rule, and who replaced them.

Public views.

- Individual page (history + diffs); downloadable .json.
- Heatmap on initiative pages showing aggregate COI concentration by sector/country.

KPIs.

- COI disclosure rate and timeliness
- Recusal compliance (flags resolved before meetings)
- Number of reassignments due to COI triggers (healthy sign the system works)

7.3 Traceable edits and authorship diffs

Purpose. Make the evolution of text and data visible to non-specialists and researchers alike.

Diff standards.

- Text diffs: paragraph-level and sentence-level views; color-coded add/remove; "semantic diff" highlighting changes in *claims* (e.g., numbers, qualifiers like "may"→"will").
- Data diffs: tables of changed rows/columns, schema alterations, new/removed datasets; summary of effect on key estimates (Δpoint estimate, ΔCI).
- Model diffs: versioned hyperparameters, training data changes, identification strategy notes.

Authorship & provenance.

- Authorship ledger: each section tagged with editor IDs, timestamps, and commit messages; link to COI records.
- Attribution badges: “substantive edit,” “clarity edit,” or “formatting only.”
- Bot disclosure: if automated tools assisted drafting, declare prompts/models and human review.

Access features.

- Side-by-side viewer for lay readers; raw patch files for experts.
- Explain change tooltip authored by the editor (≤140 chars).

Anti-tamper.

- Immutable hash journal (append-only); prior versions remain public.
- Third-party mirrors allowed (license-compatible) to deter silent edits.

KPIs.

- % of versions with full text/data/model diffs
- Share of “semantic diffs” containing quantified impact statements
- User engagement (diff page dwell time; explanations clicked)

7.4 Mandatory minority reports attached to outputs

Purpose. Preserve principled disagreement as a first-class artifact with equal permanence and discoverability.

Eligibility & scope.

- Any participant in deliberation (citizen assembly member, rapporteur, red-team lead, external commenter) may submit a minority report when they judge that: (a) core evidence is insufficient/misread, (b) ethical concerns remain unresolved, or (c) safeguards are inadequate.

Format (≤2 pages + annex).

1. Claim(s) disputed: reference claim_ids.
2. Grounds: empirical, ethical, legal, cultural/spiritual (check boxes + narrative).

3. Proposed safeguard(s) or alternative(s).
4. Test or trigger: condition under which the majority position should be revised.
5. COI & authorship disclosure.

Process & parity.

- Submission window: tied to each draft stage; late submissions allowed with justification.
- Mandatory linkage: final output must cite and link to minority reports in the executive summary and within the relevant sections ("consensus contours").
- Search & surface: minority reports appear in registry results and APIs just like main outputs; not buried in appendices.

Aftercare.

- Watch-list binding: if a minority report specifies falsifiers/triggers, they enter the monitoring plan; status lights (green/amber/red) show on the public dashboard.
- Revisit duty: when triggers fire, conveners must publish a formal reconsideration memo within a set SLA.

KPIs.

- Minority-report inclusion rate (per initiative and per draft)
- % of minority triggers integrated into monitoring plans
- Number of revisions materially influenced by minority reports (with examples)

Section takeaway.
Transparency by design is not a press release—it's an infrastructure: a public registry, a rigorous COI/recusal schema, verifiable diffing, and guaranteed minority reports. Together, these make influence traceable, disagreements legible, and decisions revisable—the core conditions for a convener like the WEF to help without ruling.

8. Roles and Separation of Functions

8.1 WEF as convener vs. analyst vs. advocate (firewalls)

Problem. Role-drift collapses convening (neutral table-setting), analysis (evidence building), and advocacy (arguing for an option), creating mistrust and "soft policy" by stealth.

Roles (clean definitions).

- Convener. Curates the room, process, and artifacts; guarantees fairness, traceability, plural inputs. No position on substantive outcomes.
- Analyst. Produces evidence ledgers, uncertainty dashboards, scenario sets; may compare options but does not recommend.
- Advocate. Proposes and argues for a specific option; authors position memos; must disclose interests and accept red-team review.

Firewalls (organizational + procedural).

- Org chart. Separate reporting lines: *Forum Convening Office (FCO)*, *Evidence & Methods Unit (EMU)*, *Initiative Proponents (IPs)*. No dual-hatting above manager level.
- Document headers. Every artifact carries a role stamp: [CONVENER], [ANALYST], or [ADVOCATE].
- Meeting hygiene. Chairs (FCO) moderate; analysts (EMU) present facts; advocates (IPs) get equal time; ethics rapporteurs and citizen panellists cross-examine.
- Conflict protocol. If a convener or analyst wishes to advocate, they must recuse from their original role for that initiative and hand off work to a peer.

Checks & KPIs.

- % of artifacts correctly role-stamped
- Number of recusals for role switches (healthy if >0 when advocacy occurs)
- External audits confirming firewall compliance

8.2 Independent ombuds and audit committee (publish-or-perish)

Purpose. Give the public an avenue to challenge process failures and force publication of required artifacts.

Institutions.

- Ombuds Office (OO). Receives complaints (coercion, opacity, bias), protects whistleblowers, orders remedies.
- Public Audit Committee (PAC). Independent, with cross-sector/region membership; oversees the registry, COI/recusal compliance, and SLAs.

Mandates.

- Publish-or-perish rule. No initiative may proceed to outreach or “launch” until mandatory artifacts are live in the registry; PAC can halt promotion and require corrections.
- Randomized audits. Quarterly sampling of initiatives for deep review (artifacts, minutes, diff integrity, COI triggers).
- Public rulings. All OO and PAC decisions are posted with timelines and implementation status.

Process flow.

1. Complaint or audit finding logged → 2) Time-boxed investigation (e.g., 30 days) → 3) Binding order: publish, correct, recuse, or pause → 4) Compliance check and public close-out.

Protections.

- Whistleblower anonymity; anti-retaliation policy with sanctions.
- Right of reply for named parties; all replies published alongside rulings.

KPIs.

- Median time to resolve cases
- % of halted initiatives due to missing artifacts (should trend to zero)
- Recurrence rate of the same violation type across initiatives

---

8.3 Ethics rapporteurs (cross-tradition) and data stewards

Purpose. Ensure moral reasoning is authentic and data are reliable, documented, and reproducible.

Ethics rapporteurs (ERs).

- Composition. Rotating pool drawn from major traditions (Abrahamic, Dharmic, human-rights jurisprudence, virtue-ethics scholarship, secular pluralists), including community practitioners—not only academics.
- Duties. Fill the ethics matrix row(s); run respectful theological/cultural reviews; write minority/majority notes; monitor safeguard uptake.

- Independence. Contracted by the PAC, not by advocates; disclose COIs; cannot be sole authors of advocacy documents.

Data stewards (DSs).

- Composition. Methodologists, statisticians, reproducibility engineers.
- Duties. Maintain evidence ledgers; verify provenance; enforce uncertainty standards; package repro bundles; run calibration checks; sign off on data/model diffs.
- Guardrails. DSs cannot author advocacy claims; they certify methods, not policies.

Interfaces.

- ERs ↔ DSs joint sign-off that the ethics matrix reflects the evidence bounds (no moral claims beyond empirical scope).
- ERs ↔ FCO ensure citizen assemblies receive balanced ethical primers.
- DSs ↔ EMU coordinate on containers/seeds for reproduction.

KPIs.

- % of initiatives with named ERs and DSs (and public bios/COIs)
- Reproduction pass rate within 90 days
- Share of proposals revised after ER safeguard requests

8.4 Escalation paths when principles clash

Problem. Pillars can conflict (e.g., transparency vs. privacy; subsidiarity vs. cross-border harms). Without a route to resolve clashes, process stalls or gets gamed.

Three-step escalation.

1. Structured trade-off memo (STM). Authors fill a one-page matrix: which principles clash, stakes, affected groups, options A/B/C with pros/cons, and proposed safeguards. Include ethics and data sign-offs.
2. Dual review.
    - Procedural review by FCO/PAC: does the STM meet process quality?
    - Substantive review by a *Principles Council* (PC) comprising ERs, DSs, citizen assembly delegates, and an external legal expert.

3. Decision & publication. PC issues a logged decision: choose A/B/C, or order redesign/pilot; attach minority views; set re-review triggers.

Default rules (if time-critical).

- Harm-minimization default. In emergencies, allow narrowly tailored action with heightened uncertainty disclosure, strict timeboxes, and automatic sunset.
- Reversibility bias. Prefer options that are easier to undo when evidence improves.

Triggers for mandatory escalation.

- Fundamental rights implicated (speech, conscience, bodily integrity)
- Indigenous lands/FPIC issues
- Non-trivial cross-border externalities
- Significant dissent from ERs or citizen assemblies

KPIs.

- Number of escalations per year (should reflect judicious use, not avoidance)
- Time from STM submission to PC decision
- Post-decision review outcomes (were triggers hit; were reversals manageable?)

Section takeaway.
Clear role boundaries, empowered independent oversight, designated ethics and data custodians, and a principled escalation path keep the machinery honest. The WEF can convene vigorously, analyze rigorously, and host advocacy—without letting any one function capture the others or override the public's right to decide.

9. Metrics and Verification

9.1 Legitimacy KPIs: participation breadth, dissent inclusion, reversal rates

Purpose. Verify that processes are representative, contestable, and corrigible.

KPI set (define, measure, target ranges).

- Participation breadth (PB).
  PB = unique_participants / eligible_population_proxy.

Cut by stratum (region, income, sector, faith/indigeneity).
Target: PB_strata within +/- 20% of population weights, or documented mitigation.

- Speaker diversity index (SDI).
  SDI = 1 - sum_i p_i^2 over speaker-origin strata (Herfindahl form).
  Target: SDI >= 0.75 for plenaries; >= 0.65 for working groups.

- Affected-stakeholder ratio (ASR).
  ASR = affected_community_participants / total_participants.
  Target: ASR >= 0.33 in design phases; >= 0.50 for pilots touching rights.

- Dissent inclusion rate (DIR).
  DIR = minority_memos_attached / minority_memos_submitted.
  Target: DIR = 1.00 (all memos published); response_log coverage = 100%.

- Dissent impact rate (DIr).
  DIr = memos_causing_material_change / minority_memos_submitted.
  Target: context dependent; track trend; avoid 0% (red flag) and 100% (rubber-stamping).

- Reversal rate (RR).
  RR = count(recommendation_revisions_or_withdrawals) / total_recommendations per year.
  Normative read: non-zero is healthy; spikes require root-cause analysis (data shifts vs. initial overreach).

- Transparency SLA adherence (TSLA).
  TSLA = artifacts_published_on_time / artifacts_due.
  Target: >= 0.95.

Data sources. Registry analytics, sign-up rosters, stratified sampling frames, response logs.

9.2 Quality KPIs: forecast calibration, replication, error-bar accuracy

Purpose. Ensure claims track reality and are reproducible.

Forecast calibration.

- Brier score (BS). For binary forecasts p_t vs. outcome o_t in {0,1}:
  BS = mean_t (p_t - o_t)^2. Lower is better.
  Target: BS improves over rolling 12 months; compare to naive baselines.

- Reliability slope/intercept. From calibration curve of predicted vs. empirical probabilities.
  Target: slope ~ 1.0, intercept ~ 0.0; report CIs.

- Log score (LS). For probabilistic statements: LS = mean_t log P(o_t).
  Target: higher is better; track vs. benchmarks.

Replication & robustness.

- Independent replication rate (IRR_90).
  IRR_90 = replications_confirmed_within_90d / claims_with_repro_bundle.
  Target: >= 0.80; if proprietary data, use surrogate tests with stated limits.
- Spec sensitivity pass (SSP).
  SSP = claims whose sign/magnitude remain within predeclared bounds under alt-specs / claims tested.
  Target: >= 0.85.
- Drift detection latency (DDL).
  DDL = mean days from performance drift trigger to public update.
  Target: <= 30 days.

Error-bar accuracy.

- Interval hit rate (IHR_95).
  IHR_95 = share of realized outcomes that fall within the published 95% interval.
  Target: 0.90–0.98 (avoid under- and over-confident extremes).
- Width efficiency (WE).
  WE = mean interval_width / outcome_scale.
  Target: narrower at equal IHR over time (learning).

Data sources. Evidence ledgers, uncertainty dashboards, post-hoc outcome datasets, replication reports.

9.3 Equity & dignity indices: distributional and cultural checks

Purpose. Detect who pays, who benefits, and whether dignity/conscience are respected.

Distributional checks.

- Benefit/cost equity index (BCEI).
  Compute benefits_t and costs_t by stratum s; report Gini or Theil T for net_benefit_s.
  Target: decreasing inequality over iterations, or compensatory mechanisms documented.

- Loser floor guarantee (LFG).
  LFG = min_s net_benefit_s relative to baseline.
  Target: LFG >= 0 for protected strata after transfers/mitigations.

Access & agency.

- Opt-out usability rate (OUR).
  OUR = successful_opt_outs_processed / opt_out_attempts.
  Target: >= 0.95; median processing time <= 7 days.
- Conscience accommodation rate (CAR).
  CAR = conscientious_objection_cases_resolved / cases_filed.
  Target: 100%; SLA compliance >= 0.95; zero retaliation findings.

Dignity/cultural metrics (from Section 5.4 rubric).

- Dignity score (DS 0..3). Mean of rubric item 1 across reviewers; publish dispersion.
- Conscience space score (CSS 0..3). From rubric item 2.
- Cultural continuity score (CCS 0..3). From rubric item 4.
  Target: DS, CSS, CCS >= 2.0 before scale-up; if < 2.0, require safeguards or limit scope.

Indigenous & FPIC.

- FPIC compliance (FPICc).
  FPICc = communities_with_valid_FPIC / communities_affected.
  Target: 1.00 where applicable.

Data sources. Impact briefs, compensation ledgers, anonymized case logs, independent cultural/theological reviews.

9.4 Post-implementation scorecards and public retrospectives

Purpose. Close the loop: compare promises to outcomes, publish what changed, and decide whether to continue, modify, or sunset.

Scorecard (one page per initiative).

- Headline verdict. Continue / Modify / Pause / Sunset (with rationale).
- Promise vs. outcome table. Each KPI with: promised target, realized value, delta, root-cause note.

- Safeguards status. Which were triggered, added, or removed; minority-trigger lights (green/amber/red).
- Equity & dignity summary. BCEI, LFG, DS/CSS/CCS, FPICc, grievance stats.
- Calibration panel. BS, IHR_95, WE, IRR_90.
- Costs & benefits. Economic, social, cultural; include non-monetizable benefits with citations.
- Voices. 3–5 short quotes from citizen assembly members, affected communities, and dissent authors.

Cadence.

- Initial review: 90 days after pilot start.
- Periodic: quarterly during pilots; semiannual in scale; mandatory pre-sunset review.

Retrospective format (public).

- 30–60 min video briefing; slides and transcript posted.
- Open Q&A window (7–14 days) with response log.
- Repository bundle: data, code diffs, ethics matrix updates, audit notes.

Decision rules (tie to falsifiers).

- If any pre-committed falsifier is tripped, default to Pause pending redesign.
- If IHR_95 < 0.8 or BS worsens vs. baseline for 2 consecutive periods, trigger methodological review.
- If LFG < 0 for protected strata, require immediate compensatory action or narrowed scope.

Governance hooks.

- PAC sign-off required for Continue in the presence of active minority red flags.
- Ombuds can compel additional safeguards or order a new citizen assembly.

KPIs (meta).

- On-time scorecard publication rate
- % of initiatives adjusted or sunset via scorecard evidence

- Public engagement: unique views, Q&A submissions, follow-up actions logged

---

Section takeaway.
Legitimacy, quality, equity/dignity, and disciplined retrospectives form a closed verification loop. With clear formulas, targets, and public scorecards, conveners like the WEF can demonstrate not just activity but accountable learning—and earn permission to keep helping without deciding.

10. Case Studies (Templates)

10.1 AI safety standard under this architecture

Use case. A voluntary model-evaluation standard for foundation models (robustness, misuse resistance, transparency).

Scope. Applies to model developers and deployers; opt-in; no ties to market access or procurement.

Roles.
[CONVENER] Forum Convening Office (FCO) — runs process
[ANALYST] Evidence & Methods Unit (EMU) — designs tests, stats, calibration
[ADVOCATES] Industry/workshop proponents — propose specific tests/thresholds
[DS] Data Stewards — verify ledgers, repro bundles
[ER] Ethics Rapporteurs — populate ethics matrix, review dignity risks
[PAC/OO] Audit & Ombuds — enforce publish-or-perish, adjudicate complaints

Process timeline (indicative).

- Month 0–1: Scoping note, agenda, invite list; COIs posted.
- Month 2: Draft v0.1 test suite + Evidence Ledger (data provenance, scoring), Uncertainty Dashboard prototype.
- Month 3: Citizen Assembly (sortition; primers on AI harms/benefits; outputs with minority views).
- Month 3–4: Open Comment on v0.5; Red-Team exercise (attack trees: jailbreaking, data extraction, capability overclaim, eval gaming).
- Month 5: v0.9 near-final with Response Log; Ethics Matrix and cultural/spiritual impact brief published.

- Month 6: Adoption Contract template released; Pilot phase (3–6 months), automatic Sunset unless re-adopted.

Core artifacts.

- Test catalog (e.g., adversarial robustness, prompt-injection resistance, eval reproducibility, hazard-capability containment) with CI/credible intervals on scores.
- Falsification Sheet: e.g., "If two independent labs cannot reproduce robustness deltas ≥ 70% of claimed within 60 days → downgrade to 'pilot-only'."
- Ethics Matrix (abbrev.):
    - *Human rights:* Cond — supports safety, ensure due process for affected users.
    - *Capabilities:* Y — preserves agency; risk: chilling effects if misused.
    - *Virtue ethics:* Cond — prudence/temperance; guard against performative signaling.
    - *Abrahamic/Dharmic/Secular pluralism:* Cond/Y with conscience/consent safeguards.

Safeguards.

- Non-coercion clause in Adoption Contracts (standards ≠ gate for essential services).
- FPIC-style review where biometric or sacred-domain models are implicated.
- Mandatory Minority Reports attached to standard (e.g., critics of capability benchmarks).
- Reversibility: pilots only; sunset after 12 months unless re-consented.

KPIs (starter set).

- IRR_90 (replication rate of eval results) ≥ 0.8
- Interval Hit Rate IHR_95 between 0.90–0.98
- DIR = 1.00 (all minority memos published), DIr tracked
- Opt-in adoption across ≥ 3 regions with at least one non-Western lab

Post-implementation scorecard.

- Promise vs. outcome (calibration, drift incidents, mitigations); quotes from citizen panel; list of minority triggers and status.

10.2 Climate finance mechanism with ethics matrix + citizen assembly

Use case. A voluntary cross-border carbon adjustment (CBCA) with LDC exemptions, revenue recycling, and technical assistance for exporters.

Scope. Participating jurisdictions opt-in; aligns with WTO principles; explicit conscientious-objection and indigenous safeguards.

Roles.
[CONVENER] FCO; [ANALYST] EMU (distributional analysis, leakage modeling); [ADVOCATES] coalition of member states/NGOs; [ER] faith/culture rapporteurs; [DS] data lineage for emissions factors and trade flows; [PAC/OO] oversight.

Process timeline (indicative).

- Month 0–1: Scoping; competence tests (local vs. cross-border locus); invite impacted exporters, indigenous groups.
- Month 2: Draft v0.1 design; Evidence Ledger for emissions factors, elasticity estimates, leakage assumptions.
- Month 2–3: Ethics Matrix populated (rights, capabilities, virtues, Abrahamic, Dharmic, secular pluralism).
- Month 3: Citizen Assembly in each participating region (localized primers; translators; travel/childcare provided).
- Month 3–4: Open Comment with Response Log; Red-Team on trade diversion, unjust weights/measures, rent-seeking, LDC harm.
- Month 5: v0.9 with safeguard pack; Adoption Contracts and Pilot lanes (narrow product sets).
- Month 6+: Pilot launch; Uncertainty Dashboard (price bands, distributional impacts, worst-plausible scenarios). Sunset at 12 months unless re-adopted.

Ethics Matrix (abbrev. fill-ins).

- *Human rights:* Cond — life/health benefits if due process and non-discrimination held.
- *Capabilities:* Y — net capability gains with targeted transfers.
- *Virtue ethics:* Cond — prudence/justice if exemptions/mercy for the least.

- *Abrahamic:* Cond — stewardship/justice; avoid “unjust weights”; require alms-like adaptation fund.
- *Dharmic:* Y — ahimsa; compassion pathways for workers.
- *Secular pluralism:* Cond — opt-in, reversibility, WTO-consistency.

Safeguards.

- LDC floor: exemption + technical assistance track.
- Distributional backstop: automatic transfer if LFG < 0 for protected strata.
- Indigenous FPIC: mandatory, with veto on projects touching sacred lands.
- Double-key trigger: expansions require both citizen assembly approval and ER green light.

KPIs.

- BCEI trending toward equity; LFG ≥ 0 in pilot; FPICc = 1.00 where applicable.
- Calibration of emissions factors (IHR_95 in target range); leakage error ≤ pre-set bound.
- DIR = 1.00; % of citizen assembly conditions integrated ≥ 60%.

Post-implementation scorecard.

- Promise vs. outcome (emissions, prices, equity, dignity scores); minority triggers dashboard; decision (continue/modify/pause/sunset).

10.3 Public health guidance with uncertainty dashboards + minority report

Use case. Voluntary guidance for a respiratory-virus surge (ventilation targets, masking in specified contexts, sick-leave norms).

Scope. Not a mandate; jurisdictions, schools, workplaces may adopt via Adoption Contracts with local tailoring and opt-outs; sacred-space carve-outs by default.

Roles.
[CONVENER] FCO; [ANALYST] EMU (epi models, causal identification notes); [ADVOCATES] professional societies; [ER] Abrahamic/Dharmic/rights counsel; [DS] data lineage for surveillance; [PAC/OO] oversight.

Process timeline (indicative).

- Week 0–1: Rapid scoping; Harm-minimization default allowed (time-boxed), with strict uncertainty disclosure.
- Week 1–2: Draft v0.5 with Evidence Ledger (surveillance biases, ascertainment rates, ventilation effect sizes, masking RCT/meta-analytic intervals), Uncertainty Dashboard, and pre-committed Falsifiers (e.g., hospitalization thresholds, calibration checks).
- Week 2: Citizen Mini-Assembly (short format) for affected groups (teachers, elder-care workers, faith leaders).
- Week 2–3: Open Comment + Response Log; Red-Team on unintended harms (stigma, school access inequities, sacred-space conflicts).
- Week 3: v0.9 plus Minority Reports (e.g., critics disputing masking efficacy in low-prevalence contexts).
- Week 4: Pilot in volunteer districts; automatic Sunset at 8 weeks unless re-adopted.

Uncertainty Dashboard (illustrative tiles).

- Point estimates with 50/95% CIs for ventilation ACH targets vs. infection risk; masking effect under varying prevalence; absenteeism baselines.
- Sensitivity to priors and measurement bias; reliability plots vs. hospitalizations.
- External-validity map (indoor density, climate, building stock).
- Dissent summary: top two minority claims + links.
- Falsifier panel: current status (green/amber/red).

Safeguards.

- Sacred-space default opt-out unless community consents.
- Dignity protections: no compelled signage that stigmatizes; plain-clothes support staff; complaint redress SLA.
- School equity fund to avoid burdening low-income families (e.g., for HEPA units, sick-leave coverage).
- Reversibility: clear exit ramps; weekly review against falsifiers.

KPIs.

- IHR_95 for hospitalization forecast bands in target; Brier score improved vs. baseline.

- OUR ≥ 0.95 (opt-out processing); CAR = 100% (conscience cases resolved, no retaliation).
- DIR = 1.00; % of minority triggers monitored = 100%; time-to-update after triggers ≤ 7 days.

Post-implementation scorecard.

- Promise vs. outcome (transmission proxies, attendance, dignity/cultural incidents); testimonials from assembly members; continuation decision with rationale tied to falsifiers.

Section takeaway.
These templates show the five pillars *at work*: evidence ledgers and uncertainty dashboards (humility), ethics matrices and theological reviews (pluralism), citizen assemblies, red-teams, and open comments (guardrails), adoption contracts, pilots, sunsets (consent), and registries/COI/diffs/minority reports (transparency). They're plug-and-play: swap in your topic, keep the interfaces.

11. Implementation Roadmap for the WEF

11.1 First 90 days: artifact registry, COI schema, red-team roster

Objective. Stand up the minimum infrastructure so no initiative can move without traceability, integrity checks, and adversarial review capacity.

Milestones (by 30/60/90 days).

- Day 30
    - Artifact Registry (alpha). Public landing page + API stub; fields for initiative meta, agendas, attendees, minutes, draft ladder, funding, COI links, and integrity hashes.
    - COI Schema (v0.9). Machine-readable JSON spec; attestation form; update cadence policy (quarterly or before relevant meetings).
    - Red-Team Charter (v0.9). Scope (technical, legal, cultural/spiritual), independence rules, severity scoring, SLAs.
    - Role Firewalls posted. Org chart separating [CONVENER] FCO, [ANALYST] EMU, [ADVOCATE] IPs.

- Day 60
    - Registry (beta). Full lifecycle fields (ethics matrix, uncertainty dashboards, adoption contracts, minority reports, scorecards); publish-or-perish gate wired.
    - COI Directory online. Personal pages with history/diffs; auto-flag triggers and recusal logging live.
    - Red-Team Roster. 30–50 vetted reviewers across domains/regions; conflict-screened; contract templates signed.
    - Ombuds + Public Audit Committee (PAC) seated. Case intake portal; randomized audit protocol approved.
- Day 90
    - Evidence Ledger kit. Templates + containers/notebooks; “no ledger, no publication” rule active.
    - Uncertainty Dashboard generator. One-page layout, CI/cred intervals, calibration tiles, dissent and falsifier panels.
    - Citizen Assembly kit (lite). Sampling scripts, primer template, facilitation rubric, accessibility supports.
    - Ethics Rapporteur pool. Named experts/practitioners across traditions with bios/COIs; review SOP published.

Owners & RACI.

- FCO: registry governance, meeting hygiene, citizen-assembly ops (R).
- EMU/DS: evidence ledgers, uncertainty dashboards, repro bundles (R).
- PAC/OO: audits, enforcement, complaint resolution (A).
- IPs: supply drafts, respond to comments, accept red-team exercises (C/R).
- ERs: ethics matrix rows, theological/cultural reviews (C/R).

KPIs (90-day check).

- Registry completeness ≥ 80% on live initiatives; TSLA ≥ 0.95
- COI disclosure rate ≥ 95%; recusal compliance = 100% on flagged cases
- Red-team roster coverage across at least 5 regions and 8 disciplines

11.2 Six-month pilots: one initiative per pillar

Objective. Prove each pillar in the wild with a bounded, real initiative; prioritize reversibility and measurement.

Pilot slate (illustrative; each aligned to a pillar).

1. Humility pilot: AI evaluation standard (Section 10.1) — stress test evidence ledgers, falsifiers, uncertainty dashboards.
2. Pluralism pilot: Climate finance CBCA module (Section 10.2) — run ethics matrix, faith/culture reviews across 3 regions.
3. Guardrails pilot: Digital identity UX guideline — citizen assemblies, open drafting, red-team on coercion risks.
4. Consent pilot: Public health guidance (Section 10.3) — adoption contracts, pilots, sunset provisions, opt-out SLAs.
5. Transparency pilot: End-to-end registry exercise — traceable diffs, COI/recusal, minority reports with parity.

Common pilot rules.

- Scope: smallest responsible slice; time-boxed (8–16 weeks).
- Contract: plain-language adoption contract with exit ramps; non-adoption neutrality clause.
- Monitoring: weekly dashboard updates; public changelogs.
- Scorecards: 30-day and end-of-pilot scorecards with “continue/modify/pause/sunset” verdict.

Cadence & gates.

- Month 1–2: pilots 1–2 launch; Month 3–4: pilots 3–4; Month 5–6: transparency pilot overlays all.
- Go/No-Go gate: any missing artifacts → NO-GO; any falsifier tripped → PAUSE unless PAC approves narrowly tailored continuation.

Six-month KPIs.

- ≥ 4 pilots completed; ≥ 2 outside Europe/N. America
- DIR = 1.00 (all minority memos published); ≥ 60% of substantive comments acted on

- RR (reversal rate) non-zero with documented causes; IHR_95 within target on pilots with forecasts
- BCEI improving or compensations documented; FPICc = 1.00 where applicable

---

11.3 Twelve-month scale: annual legitimacy report and APIs for partners

Objective. Institutionalize the model, publish what changed, and make the system interoperable.

Deliverables.

- Annual Legitimacy Report (ALR).
    - Sections: Participation breadth, dissent inclusion/impact, reversals; calibration and replication stats; dignity/cultural outcomes; registry uptime/completeness; case studies; remedial actions.
    - Audit letter from PAC; ombuds summary of complaints and remedies.
- Public APIs (v1).
    - initiatives, drafts, diffs, funding, coi, minority_reports, scorecards.
    - Rate-limited keys for media/academia; bulk export for watchdog mirrors.
- Partner Tooling.
    - Ethics Matrix Builder (guided prompts + citation slots); Citizen Assembly Kit (sampling & primer generator); Uncertainty Dashboard SDK (chart functions, schema validation).
- Training & accreditation.
    - Convening facilitators, data stewards, ethics rapporteurs, red-team leads—short courses with open syllabi and public rosters of accredited practitioners.

Policy codification.

- Hard-code publish-or-perish into WEF initiative SOPs; embed firewalls in HR/performance plans; require adoption contracts for any field pilots.

12-month KPIs.

- ALR published on time; independent replication of ALR metrics by at least one external group

- API uptime ≥ 99.5%; ≥ 10 partner orgs consuming APIs; ≥ 3 watchdog mirrors online
- ≥ 50% of new initiatives launched under the five-pillar architecture; TSLA ≥ 0.97
- Public satisfaction proxy: increase in unique registry users and media citations that include uncertainty intervals

---

11.4 Interoperability with governments, NGOs, faith bodies, academia

Objective. Make the architecture portable; let others plug in while preserving local authority and consent.

Interfaces & agreements.

- Data & artifact standards. Publish JSON schemas for evidence ledgers, ethics matrices, response logs, scorecards; align where possible with open government data standards and research reproducibility norms.
- MOUs with public authorities.
    - WEF commits to convening/analysis firewalls and transparency SLAs.
    - Authorities commit to keep adoption opt-in and to run competence tests for locus decisions.
    - Joint escalation rules (rights, FPIC, cross-border harms) and re-consent triggers.
- Faith & cultural councils. Standing MOUs with interfaith bodies and indigenous organizations for rapid theological/cultural review; remuneration and timeline guarantees; consent for publication of reasoned opinions.
- Academic partnerships. Methods clinics with universities for replication, calibration audits, and red-team research; shared datasets with privacy protections; co-authored retrospectives.

Technical interoperability.

- APIs for ingestion. Allow governments/NGOs to push artifacts (assemblies, comments, minority reports) into the registry with their own namespaces and governance metadata.
- Identity & access. Federated identity (OIDC/SAML) for authenticated contributors; public read by default.

- Localization. Multilingual artifacts; right-to-left support; region-specific ethics-matrix presets.

Safeguards for sovereignty & consent.

- No exclusive rights or proprietary lock-ins; permissive licenses for templates/tools.
- Explicit clauses that WEF artifacts are advisory; governmental adoption requires domestic due process.
- FPIC enforcement where indigenous rights are implicated; conscientious-objection pathways embedded.

Interoperability KPIs.

- Number of external bodies publishing artifacts to the registry
- % of partner-submitted artifacts passing schema validation on first try
- Turnaround time for interfaith/indigenous reviews; share integrated into final designs
- Instances where governmental due process cites and adapts—rather than copies—WEF artifacts (evidence of local sovereignty)

Section takeaway.
In 90 days, the plumbing; in 6 months, proof; in 12 months, institutionalization and APIs. With clear owners, gates, and KPIs—and real interoperability—the WEF can operationalize “help, don’t decide” as a living system, not a slogan.

12. Risks and Antipatterns

12.1 Performative transparency

What it is. Publishing lots of information with little *use*: artifacts that are incomplete, unsearchable, late, or so technical that outsiders can’t act on them.

Failure modes.

- Data dumps without levers. Agendas/minutes posted but without links to drafts, response logs, or decision hooks.
- Opacity by complexity. PDFs with screenshots of tables (no machine-readable data); missing COI fields; broken permalinks.

- Asymmetric visibility. Glossy press pages while red-team reports/minority memos are buried or delayed.

Leading indicators.

- Registry completeness score < 0.8; TSLA (on-time publication) < 0.95.
- Low API usage; high bounce rate on artifact pages; few external citations to uncertainty intervals.
- “Consensus” statements that lack links to dissent or falsifiers.

Countermeasures.

- Publish-or-perish gate: no promotion until mandatory artifacts are live *and* machine-readable.
- One-page “How to Act” card per initiative (how to comment, request data, file a conflict).
- Diff & dashboard by default: every version bump ships with a human-readable changelog and the uncertainty dashboard on the landing page.
- Usability audits: external testers rate findability, readability, and machine-readability quarterly.

KPIs to watch.

- API uptime/consumption; % artifacts with JSON alongside PDF; median time to fix broken links; number of media/academic pieces citing intervals (not just point claims).

12.2 Token pluralism

What it is. Inviting diverse voices but giving them no real influence; ethics matrices filled perfunctorily; citizen panels used as optics.

Failure modes.

- Checkbox representation. One-off faith/culture consults with no guardrail adoption.
- Flattened disagreement. Ethics matrix rows say “conditional” without recording non-negotiables or opt-out clauses.
- Unbalanced agenda time. Critics receive short slots versus incumbents’ keynotes.

Leading indicators.

- Dissent impact rate (DIr) ~ 0% over time; % of citizen assembly conditions adopted < 30%.
- Minority reports exist but are uncited in the executive summaries; safeguard clauses absent.
- Speaker diversity index looks good, but authorship of final text is homogenous.

Countermeasures.

- Parity rules. Minority reports must be cited in the *executive summary* and tracked on dashboards.
- Double-key triggers. High-risk provisions require both citizen assembly and ethics council approval.
- Guardrail budget. Allocate funds/time for implementing pluralism-driven safeguards (opt-outs, FPIC, compensation).
- Authorship audits. Publish authorship diffs and sector/geography of contributors to the final text.

KPIs to watch.

- Share of safeguards originating from dissent that make it into the adoption contract; ratio of talk-time vs. text-time (who spoke *and* who wrote); post-hoc surveys from invited critics (did input matter?).

12.3 Incumbent capture

What it is. Well-resourced actors steer agenda, metrics, and timelines to entrench advantage—often via standard-setting, certification monopolies, or procurement gating.

Failure modes.

- Standards as moats. Technical criteria that only incumbents can meet (IP-locked test suites; proprietary data dependencies).
- Rotating-door authorship. Vendors drafting the rules they will later certify/enforce.
- Backdoor coercion. Ratings agencies and financiers encode “voluntary” standards into eligibility criteria.

Leading indicators.

- High COI heatmap concentration in a few sectors/countries; repeated authors from the same firms; red-team findings about market lock-in ignored.
- Adoption skew: rapid uptake by large players; low or zero uptake by SMEs or low-income regions.
- Procurement or trade groups start requiring the standard absent due process.

Countermeasures.

- Firewall enforcement. Analysts/organizers cannot advocate; vendors that advocate must recuse from drafting; PAC can pause initiatives for COI breaches.
- Open test suites. Reference implementations and datasets under permissive licenses; SME pathways; affordability caps on certification.
- Anti-tying clauses. Adoption contracts forbid linking the standard to unrelated market access; trigger coercion audits.
- Diversity quotas for authorship. Minimum SME/Global South co-authors for any draft moving to v0.9.

KPIs to watch.

- SME adoption rate; Global South participation; number/time to resolve COI-triggered recusals; price of compliance relative to revenue for small actors; instances of procurement/rating gatekeeping reversed.

12.4 Consultation fatigue (rotate formats; compensate civic labor)

What it is. Over-asking communities and civil society to engage, with little feedback loop or compensation—leading to lower quality input and eroded trust.

Failure modes.

- Serial listening with no action. Many workshops; few response logs or visible changes.
- Barrier-heavy asks. Long weekday sessions, English-only materials, no childcare or travel support.
- Unpaid expert labor. Repeated requests to community leaders and theologians without honoraria.

Leading indicators.

- Declining response rates; rising dropout mid-assembly; skew toward professional NGO voices.
- Negative satisfaction in post-session surveys; repeat commenters asking “what changed?”
- Geographic or demographic participation gaps widen over time.

Countermeasures.

- Rotate formats. Mix mini-assemblies, asynchronous comment portals, SMS polls, and neighborhood workshops; cap participation time per invitee.
- Compensate civic labor. Honoraria for citizen panellists; paid time for community/faith leaders; micro-grants for local convenings.
- Service design. Evening/weekend options; childcare, accessible venues, multilingual primers; travel stipends.
- Tight feedback loops. Publish 1–2 page “You said / We did” after each phase; show diffs tied to specific comments.

KPIs to watch.

- Participation breadth over time (by stratum); median time burden per participant; satisfaction delta (pre/post); % comments with individualized responses; honoraria coverage rate for non-salaried contributors.

Section takeaway.
These antipatterns—performative transparency, token pluralism, incumbent capture, and consultation fatigue—are predictable. Treat them as monitored risks with concrete indicators and pre-built countermeasures. If the dials start drifting, you don’t need a scandal to course-correct—you’ve already wired the fix into the system.

13. Ethical/Theological Appendix: Hosting Disagreement Well

13.1 Minimal common ground (dignity, non-coercion, truthfulness)

Purpose. Establish a thin, cross-tradition baseline that enables collaboration without forcing doctrinal agreement.

Three axioms (plain language).

1. Dignity. Every person has inherent worth; no one is used merely as a means.
2. Non-coercion. Consent is meaningful only when exit is real and retaliation absent.
3. Truthfulness. Claims, evidence, and motives are presented honestly; errors are corrected openly.

Operationalization.

- Dignity checks embedded in the impact rubric (Section 5.4): screen for stigmatizing categories, compelled expression, degrading treatment; require respectful UX and language guidelines.
- Non-coercion guarantees in every Adoption Contract (Section 6.2): bold “no penalty for non-adoption” clause; opt-out SLAs; coercion audits for tying/gatekeeping.
- Truthfulness norms: evidence ledger completeness; uncertainty dashboards on first pages; documented corrections with versioned diffs.

Meeting norms (short rule-set).

- Name arguments, not persons.
- Disclose material interests up front (link to COI).
- Charitable restatement: before critique, summarize the other view to their satisfaction.
- Time parity: equal speaking time for minority and majority views in deliberation windows.

Metrics.

- Dignity score ≥ 2.0/3.0 before scale-up.
- Opt-out usability rate ≥ 95%.
- Correction latency ≤ 14 days for factual errors.

13.2 Conscientious objection and safe opt-outs

Purpose. Protect moral agency when proposals conflict with sincerely held beliefs or identity-defining practices.

Who qualifies. Individuals or communities (including faith bodies and indigenous groups) asserting a specific, reasoned moral/theological ground for non-participation, with minimal documentation burden.

Process (simple, time-boxed).

1. Notice. Submit a one-page form: objection basis, scope (which parts of the initiative), and proposed alternative/compliance-neutral pathway.
2. Immediate shield. Provisional protection kicks in upon receipt—no penalties, no shaming UX.
3. Dialogue. 7–14 day structured conversation with facilitators to clarify scope and explore accommodations.
4. Decision. Grant accommodation, provide an equivalent alternative, or agree to respectful separation from the initiative's scope.
5. Record. Publish anonymized decision + rationale in the registry; update Adoption Contract if systemic.

Design features.

- Equivalence, not second-class. Alternatives must deliver comparable access to services/benefits without stigma cues.
- Privacy. Public logs redact personally identifying or sensitive doctrinal details unless consented.
- Appeal path. Ombuds can review within 14 days; reversals published.

Example clauses (insert in Adoption Contracts).

- *"Nothing in this initiative requires speech, symbols, or acts contrary to conscience. Where conflicts arise, participants may opt out of the conflicting element and receive a functionally equivalent alternative without delay or disadvantage."*
- *"No data about a person's religious or philosophical beliefs will be required to request accommodation; voluntary disclosures are protected and minimized."*

Anti-retaliation & monitoring.

- Prohibit reputational downgrades, access throttling, or procurement penalties due to objections.
- Publish Conscience Accommodation Rate (CAR) and complaint resolutions; target 100% resolution with SLA compliance ≥ 95%.

13.3 Sacred-domain safeguards (speech, ritual, conscience)

Purpose. Prevent initiatives from intruding into domains traditions mark as sacred, unless freely invited by those communities.

Sacred domains (illustrative, non-exhaustive).

- Speech & symbols: confessions, vows, liturgy, sacred texts, names and images with special status.
- Ritual & space: worship sites, burial grounds, pilgrimages, sacraments/rites of passage.
- Conscience & pastoral care: confessional confidentiality, clerical counseling, moral instruction.
- Bodily integrity: fasting/feasting rules; clothing and modesty codes; anointing/laying of hands.

Safeguard package (default-on).

- No-entry default. Initiatives do not collect data, install devices, run trials, or require behaviors in sacred spaces or rituals unless the directly affected community issues free, prior, and informed consent (FPIC).
- Double-key rule. Any exception requires: (1) a supermajority of a local citizen assembly that includes the affected community, and (2) written assent from recognized faith/indigenous authorities, both revocable.
- Data minimization & segregation. If involvement is consented, collect only necessity-bounded data; store separately with access controls approved by the community; sunset on a short timer.
- Symbolic sensitivity review. Ethics rapporteurs assess names, imagery, and placement for unintended profanation or appropriation; propose alternates where needed.

- Time/place tailoring. Avoid sacred times (holy days, festivals, mourning periods) unless explicitly invited; provide quiet modes and signage neutrality near sacred sites.

FPIC protocol (lightweight).

1. Plain-language brief in relevant languages; risks/benefits; withdrawal rights.
2. Deliberation window sufficient for community processes (e.g., parish council, panchayat, elders).
3. Recorded consent (yes/no/conditional) with *conditions listed as enforceable safeguards*.
4. Ongoing consent (renewal checkpoints; easy withdrawal).

Red lines (hard exclusions).

- Compelled speech contradicting doctrine (e.g., forced slogans or symbols).
- Biometric tracking inside active worship spaces without FPIC.
- Interference with clergy-penitent confidentiality or pastoral records.

Enforcement & remedies.

- Sacred-domain breach = immediate pause. Ombuds triggers stop, publishes investigation timeline, and engages community leaders.
- Restorative process. Public apology, remediation plan (e.g., data deletion verified by third party), and material compensation if harm occurred.
- Future guardrails. Stronger defaults (e.g., geofencing, hardware beacons that disable sensors) before any re-entry is considered.

Monitoring.

- Registry tag for sacred-domain interactions; publish counts, outcomes, and time-to-remedy.
- Satisfaction surveys of participating communities post-engagement.

Appendix tools (templates).

- Conscience Accommodation Form (1 page): fields for scope, alternative requested, contact, privacy preferences.

- Sacred-Domain Checklist (pre-deployment): FPIC status, symbolic review done, timing conflicts checked, data minimization plan, exit/rollback ready.
- Dialogue Guide for Disagreement: prompts to surface value conflicts, “steel-man” exercises, and how to write conditions/safeguards.

Appendix takeaway.
A thin common ground (dignity, non-coercion, truthfulness), real conscientious-objection paths, and default-on sacred-domain safeguards let deep moral differences coexist with cooperation. Disagreement stays visible and respected—and coordination proceeds, when it should, by consent rather than compulsion.

14. Conclusion: Help, Don’t Decide

14.1 What the WEF owes the public

A principled posture. The WEF’s legitimate role is to convene, clarify, and coordinate—not to rule. That role carries duties:

- Epistemic duty: Publish evidence ledgers, uncertainty dashboards, and pre-committed falsifiers with every claim. Admit limits, document dissent, correct quickly.
- Moral duty: Ground proposals in plural traditions via the ethics matrix and theological/cultural review; record where frameworks agree, disagree, or conditionally assent.
- Democratic duty: Couple initiatives to citizen assemblies, open drafting, and red-team stress tests so ordinary people and critics can shape outcomes.
- Consent duty: Use adoption contracts, pilots, and sunset clauses; respect opt-outs and conscientious objection without penalty.
- Transparency duty: Maintain a public artifact registry, COI/recusal schema, and traceable diffs; attach minority reports with equal permanence.

Deliverable, not declarative: These duties are evidenced in artifacts, SLAs, and KPIs—visible, auditable, and reversible.

14.2 What publics should demand

A practical checklist for citizens, journalists, lawmakers, faith and community leaders:

- Where are the artifacts? Refuse to engage with glossy summaries absent ledgers, dashboards, drafts, and response logs.
- Who disagrees, and where's their page? Expect minority reports cited in executive summaries and tracked on monitoring dashboards.
- What would change your mind? Look for falsifiers, not just forecasts. If none exist, press pause.
- Can we opt out? Demand adoption contracts with real exit ramps, non-retaliation clauses, and equivalently resourced alternatives.
- Who benefits/who pays? Require equity & dignity scorecards, FPIC where applicable, and distributional backstops before scale-up.
- Whose fingerprints are on the text? Inspect authorship diffs and COI heatmaps; insist on recusals where interests collide.
- What happens after launch? Ask for post-implementation scorecards, public retrospectives, and documented reversals when triggers fire.

If any answer is "we can't show that," the public answer should be "then don't ship it."

14.3 From aura to auditability

The shift. For decades, the WEF (and similar conveners) have traded on aura—exclusive rooms, eminent names, ambitious narratives. This architecture replaces aura with auditability:

- From statements to structured claims (ledgers + intervals).
- From consensus theater to consensus contours (where agreement ends and dissent begins).
- From access to accountability (registries, diffs, COIs, minority parity).
- From momentum to measured learning (falsifiers, reversals, calibrated forecasts).
- From one-way convening to two-way consent (citizens decide, communities opt in or out).

Payoff. Conveners that “help, don’t decide” become trusted commons utilities: places where options are clarified, trade-offs are made explicit, safeguards are real, and coordination proceeds only as far as consent allows. The result is not weaker influence but legitimate influence—the kind that endures because it is continually earned.

One-line closing.
If the WEF wants to lead responsibly in a plural, democratic world, it must trade mystique for mechanisms—help, don’t decide, and make every ounce of influence auditable.

A. Ethics Matrix template (fillable)

Appendix A. Ethics Matrix (template)

One page; one row per framework. Use Y = endorse; N = oppose; C = conditional.

Framework	Verdict (Y/N/C)	Rationale (≤2 lines)	Conditions / Safeguards
Human rights (UDHR/ICCPR/ICESCR)			
Capability approach (Sen/Nussbaum)			
Virtue ethics (Aristotle/MacIntyre)			
Abrahamic duty/rights			
Dharmic traditions			
Secular pluralism/liberalism			

Filling rules

- Verdict: Y = endorse; N = oppose; Cond = conditional (list conditions).
- Citations: Quote doctrines/articles sparingly; prefer paraphrase + reference (e.g., UDHR Art. 7; Nicomachean Ethics II).
- Authenticity: Rapporteur signs row; discloses COI; notes interpretive school (e.g., usul al-fiqh, Thomist, Advaita).

B. Citizen Assembly brief kit (neutral primers + question set)

1) Primer (≤ 30 pages, grade-10 readability)

- 1.1 Two-page overview: What is proposed, why now, who is affected, what choices exist (including “do nothing”).
- 1.2 Evidence snapshot: 5–7 key findings with uncertainty bands and plain-language caveats.
- 1.3 Ethics snapshot: One-page mini–ethics matrix summary (where traditions converge/diverge).
- 1.4 Impact snapshot: Distributional map (who pays/benefits), dignity & cultural flags, known unknowns.
- 1.5 Process & rights: Your role; COI policy; opt-out; how dissent will be recorded.

2) Expert packets (balanced)

- Proponent memo (≤ 6 pages) and critic memo (≤ 6 pages) with identical structure: claim, evidence, uncertainties, falsifiers, safeguards.
- Disclosure page for each expert (COI, affiliations).

3) Question set (facilitator-ready)

- Understanding: What problem is this solving? What alternatives exist?
- Evidence: What would change our mind (falsifiers)? Where does uncertainty matter most?
- Fairness: Who bears costs/benefits? What backstops should be mandatory?
- Dignity/Conscience: Any sacred-domain conflicts? What opt-outs are needed?
- Consent: What should the adoption contract include (term, exit, monitoring)?
- Decision: Under which conditions (if any) should we recommend “yes,” “pilot,” or “no”?

4) Facilitation kit

- Timed agenda; speaking-time tracker; minority-statement template; accessibility checklist (childcare, translation, mobility); feedback survey (knowledge gain; perceived fairness).
-

5) Outputs

- Majority recommendation (≤ 2 pages), minority statements (≤ 1 page each), *conditions list*, and a two-page public summary ("You said / We did" follows later).

C. Red-Team playbook (threat models, checklists)

Goal: Systematically surface failure modes (technical, legal, economic, cultural/spiritual) before they happen.

1) Charter (copy/paste)

- Independence: Team contracted by audit committee; proposers have no veto.
- Scope: Break assumptions; model exploitation, capture, distributional harms, sacred-domain conflicts.
- Deliverables: Attack tree, severity table, mitigation pack, residual-risk statement.

2) Threat-model checklist (minimum)

- Abuse & gaming: Metric manipulation? Proxy incentives? Misreporting channels?
- Backdoor coercion: Ratings/procurement tying? "Voluntary" standard as market gate?
- Distributional harm: Who loses? Floors for the worst-off?
- Rights & dignity: Privacy, speech, conscience, sacred spaces?
- Institutional capture: Standards committees; certification monopolies; IP bottlenecks.
- International spillovers: Trade distortions; sovereignty; FPIC.
- Operational fragility: Single points of failure; rollout miscalibration; rollback feasibility.

3) Severity scoring

- Likelihood (L): 1–5; Impact (I): 1–5; Detectability (D): 1–5.
- Risk Priority Number (RPN) = $L \times I \times (6 - D)$ (higher = worse).
- Tag High (≥ 50), Med (25–49), Low (< 25).

4) Attack tree (template)

- Root: "Stakeholder coerced / Harm realized."
    - Branch A: Metric gaming → (nodes: proxy inflation, auditing blind spots…).
    - Branch B: Market gatekeeping → (nodes: ratings/prequal rules…).
    - Branch C: Sacred-domain breach → (nodes: data capture in worship spaces…).
    - Branch D: Rights conflict → (nodes: compelled speech, indiscriminate surveillance…).

5) Mitigation pack (for each High RPN)

- Design change: What to remove/limit.
- Safeguard: Opt-out, rate limit, transparency hook, FPIC, double-key approval.
- Monitoring: Trigger, threshold, owner, SLA to pause.
- Proof: How will we know the mitigation works (test/metric)?

6) Exercise formats

- Tabletop: scripted injects; cross-functional participants.
- Live-fire (where safe): bounty-like tests on process, not only code.
- Red vs. Blue write-ups published in registry.

D. Impact Assessment rubric (economic + cultural/spiritual)

Goal: Bring non-economic values into the same frame as cost/benefit, with public, repeatable scoring.

Rubric (score 0–3; attach notes & citations)

Dimension	0	1	2	3	Notes (who's affected? evidence?)
Economic efficiency (cost-effectiveness)	none/unknown	weak	moderate	strong	
Distributional equity (loser floors)	regressions	mixed; weak backstops	backstops exist	equitable with proof	
Dignity & respect	degrading/stigmatizing	risks w/ weak mitigations	mostly respectful	proactively dignifying	
Conscience space (opt-outs)	none	partial; hard to use	functional	easy; parity alternatives	
Cultural continuity	conflicts ignored	risks flagged	mitigations present	community-led fit	
Sacred-domain protection (FPIC)	violations	incomplete	compliant	exemplary FPIC & co-governance	
Relational cohesion (trust)	harms trust	ambiguous	maintained	strengthens trust	
Privacy & data minimization	invasive	unclear	limited & necessary	minimal-by-design	
Reversibility (exit ramps)	none	unclear	present	tested & rapid	

Method notes

- Mixed methods: quick ethnography, structured interviews, short surveys; cite sample frames.
- Reviewer dispersion: publish disagreement spread (min/mean/max; short rationales).
- Trigger thresholds: If *any* sacred-domain score ≤ 1 → pause or add safeguards; if *dignity* ≤ 1 for a protected group → redesign required.

E. COI/Recusal schema (machine-readable fields)

Goal: Standardize conflict declarations; enable auto-flagging and public traceability. The schema below can be used as JSON lines or in a relational store.

```
{
 "entity_id": "pers_01234",
 "name": "Full Name",
 "roles": ["author","panelist","chair"],

 "affiliations": [
  {"org": "Organization A", "role": "Senior Fellow", "since": "2023-02"}
 ],

 "coi": [
  {
   "type": "financial | fiduciary | advisory | familial | other",
   "counterparty": "Org/Person",
   "description": "Consulting / equity / board seat / grant / etc.",
   "amount_band": "<$5k | $5k–$50k | >$50k | in-kind",
   "period_start": "YYYY-MM",
   "period_end": "YYYY-MM or 'present'",
   "evidence_url": "https://…"
  }
 ],

 "topic_scope": ["carbon accounting","AI safety metrics"],
```

```
  "recusal_rules": [
    {
      "trigger": "topic:carbon accounting OR counterparty:Organization A",
      "applies_to": ["drafting","chairing","vote","certification"],
      "start": "2025-01-01",
      "end": "2026-01-01"
    }
  ],

  "attestations": [
    {"date": "2025-10-12", "attestor": "Full Name", "method": "digital-signature"}
  ]
}
```


Operational guidance

- Update cadence: quarterly or before relevant meetings.
- Auto-flags: if counterparty ∈ initiative funders OR topic_scope overlaps draft domain → require recusal before participation.
- Recusal log (public): who, from_what, rule, replacement, date.
- Heatmaps: aggregate by sector/country on initiative pages.

Compliance checks

- Disclosure rate ≥ 95%; recusal compliance 100% on flagged cases; time-to-update after role change ≤ 14 days.

Quick-reference: artifact checklists (copy/paste)

Ethics Matrix (A): verdicts, conditions, citations, rapporteur signatures, COIs.
Assembly Kit (B): primers, balanced memos, Q-set, facilitation plan, outputs, “You said / We did.”

Red-Team (C): charter, checklist, attack tree, severity table, mitigation pack, public write-ups.
Impact Rubric (D): scored table + notes, reviewer dispersion, trigger thresholds.
COI Schema (E): JSON on file, public page, auto-flags, recusal log, quarterly attestations.

Appendices takeaway: These templates turn principles into repeatable *interfaces*. Fill them, publish them, and you'll have the audit trail—and the social license—needed for a convener that helps without deciding.

References

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. "Concrete Problems in AI Safety." *arXiv:1606.06565*.

Brier, Glenn W. 1950. "Verification of Forecasts Expressed in Terms of Probability." *Monthly Weather Review* 78 (1): 1–3.

Dryzek, John S., and Simon Niemeyer. 2008. "Discursive Representation." *American Political Science Review* 102 (4): 481–93.

Fishkin, James S. 2018. *Democracy When the People Are Thinking: Revitalizing Our Politics Through Public Deliberation*. Oxford University Press.

Foot, Philippa. 2001. *Natural Goodness*. Oxford University Press.

Fung, Archon, Mary Graham, and David Weil. 2007. *Full Disclosure: The Perils and Promise of Transparency*. Cambridge University Press.

Gneiting, Tilmann, and Adrian E. Raftery. 2007. "Strictly Proper Scoring Rules, Prediction, and Estimation." *Journal of the American Statistical Association* 102 (477): 359–78.

Guerrero, Alexander A. 2014. "Against Elections: The Lottocratic Alternative." *Philosophy & Public Affairs* 42 (2): 135–78.

Habermas, Jürgen. 1996. *Between Facts and Norms*. MIT Press.

Hursthouse, Rosalind. 1999. *On Virtue Ethics*. Oxford University Press.

International Association for Impact Assessment (IAIA). 2015. *Social Impact Assessment: Guidance for Assessing and Managing the Social Impacts of Projects*.

International Labour Organization (ILO). 1989. *Indigenous and Tribal Peoples Convention (No. 169)*.

Kelly, Tim. 1998. "Arguing Safety — A Systematic Approach to Safety Case Management." *Safety-Critical Systems Symposium*.

Landemore, Hélène. 2020. *Open Democracy: Reinventing Popular Rule for the Twenty-First Century*. Princeton University Press.

MacIntyre, Alasdair. 1981. *After Virtue*. University of Notre Dame Press.

Nissenbaum, Helen. 2010. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press.

Noveck, Beth Simone. 2015. *Smart Citizens, Smarter State: The Technologies of Expertise and the Future of Governing*. Harvard University Press.

Nussbaum, Martha C. 2000. *Women and Human Development: The Capabilities Approach*. Cambridge University Press.

OECD. 2003. *Managing Conflict of Interest in the Public Service: OECD Guidelines*.

OECD. 2020. *Innovative Citizen Participation and New Democratic Institutions: Catching the Deliberative Wave*.

O'Neill, Onora. 2002. *A Question of Trust*. Cambridge University Press.

Peng, Roger D. 2011. "Reproducible Research in Computational Science." *Science* 334 (6060): 1226–27.

Rawls, John. 1993. *Political Liberalism*. Columbia University Press.

Sabel, Charles F., and Jonathan Zeitlin. 2008. "Learning from Difference: The New Architecture of Experimentalist Governance in the EU." *European Law Journal* 14 (3): 271–327.

Schoemaker, Paul J. H. 1995. "Scenario Planning: A Tool for Strategic Thinking." *Sloan Management Review* 36 (2): 25–40.

Sen, Amartya. 1999. *Development as Freedom*. Oxford University Press.

Sen, Amartya. 2009. *The Idea of Justice*. Harvard University Press.

Smith, Graham. 2009. *Democratic Innovations: Designing Institutions for Citizen Participation*. Cambridge University Press.

Spiegelhalter, David, Mike Pearson, and Ian Short. 2011. "Visualizing Uncertainty About the Future." *Science* 333 (6048): 1393–1400.

Spiegelhalter, David. 2017. "Risk and Uncertainty Communication." *Annual Review of Statistics and Its Application* 4: 31–60.

Stodden, Victoria, Friedrich Leisch, and Roger D. Peng, eds. 2014. *Implementing Reproducible Research*. Chapman & Hall/CRC.

Taylor, Charles. 1989. *Sources of the Self: The Making of the Modern Identity*. Harvard University Press.

United Nations. 1948. *Universal Declaration of Human Rights* (UDHR).

United Nations. 1966. *International Covenant on Civil and Political Rights* (ICCPR).

United Nations. 1966. *International Covenant on Economic, Social and Cultural Rights* (ICESCR).

United Nations. 2007. *United Nations Declaration on the Rights of Indigenous Peoples* (UNDRIP).

van der Bles, Anne Marthe, Sander van der Linden, Alexandra L. J. Freeman, James Mitchell, Ana-Lucia Galvao, Theresa E. Zavaleta, David Spiegelhalter, and Sir John T. Houghton. 2019. "Communicating Uncertainty About Facts, Numbers and Science." *Royal Society Open Science* 6 (5): 181870.

Weber, Max. 1919. "Politics as a Vocation." In *From Max Weber: Essays in Sociology*, edited by H. H. Gerth and C. Wright Mills. Oxford University Press (1946 edition).

World Bank. 2016. *Environmental and Social Standard 7: Indigenous Peoples/Sub-Saharan African Historically Underserved Traditional Local Communities (ESS7)*.

Optional add-ons (standards & safety):

IEC. 2010. *IEC 61508: Functional Safety of Electrical/Electronic/Programmable Electronic Safety-Related Systems*.

ISO. 2018. *ISO 31000: Risk Management — Guidelines*.